Indexation aléatoire et similarité inter-phrases appliquées au résumé automatique

Hai Hieu Vu

THESE / UNIVERSITE DE BRETAGNE SUD
sous le sceau de l Université Bretagne Loire

pour obtenir le titre de
DOCTEUR DE L UNIVERSITE DE BRETAGNE SUD

Mention : Informatique
Ecole doctorale SICMA

Présentée par
VU Hai Hieu

Préparée dans l'équipe EXPRESSION
Laboratoire IRISA

Indexation aléatoire et similarité inter-phrases appliquées au résumé automatique

Thèse soutenue le 29 janvier 2016

devant le jury composé de :

Pierre-François MARTEAU
Professeur, université de Bretagne Sud / directeur de thèse

Jeanne VILLANEAU
MCF, université de Bretagne Sud / co-directrice de thèse

Farida SAÏD
MCF, université de Bretagne Sud / co-directrice de thèse

Sophie ROSSET
Chercheuse, LIMSI – CNRS / rapporteuse

Emmanuel MORIN
Professeur, université de Nantes / rapporteur

Gwénolé LECORVÉ
MCF, université de Rennes 1 / examinateur

UNIVERSITE DE BRETANGE-SUD

Résumé

IRISA
EXPRESSION

Docteur en informatique

Indexation aléatoire et similarité inter-phrases appliquées au résumé automatique

par VU Hai Hieu

Face à la masse grandissante des données textuelles présentes sur le Web, le résumé automatique d'une collection de documents traitant d'un sujet particulier est devenu un champ de recherche important du Traitement Automatique des Langues. Les expérimentations décrites dans cette thèse s'inscrivent dans cette perspective. L'évaluation de la similarité sémantique entre phrases est l'élément central des travaux réalisés. Notre approche repose sur la similarité distributionnelle et une vectorisation des termes qui utilise l'encyclopédie Wikipédia comme corpus de référence. Sur la base de cette représentation, nous avons proposé, évalué et comparé plusieurs mesures de similarité textuelle ; les données de tests utilisées sont celles du défi SemEval 2014 pour la langue anglaise et des ressources que nous avons construites pour la langue française. Les bonnes performances des mesures proposées nous ont amenés à les utiliser dans une tâche de résumé multi-documents, qui met en oeuvre un algorithme de type PageRank. Le système a été évalué sur les données de DUC 2007 pour l'anglais et le corpus RPM2 pour le français. Les résultats obtenus par cette approche simple, robuste et basée sur une ressource aisément disponible dans de nombreuses langues, se sont avérés très encourageants.

Remerciements

Je tiens à remercier, en tout premier lieu, mon directeur et mes co-directeurs de thèse, Monsieur le Professeur Pierre-François MARTEAU, Mesdames Jeanne VILLANEAU et Farida SAÏD pour m'avoir accueilli, guidé et mis dans les meilleures conditions pour préparer ma thèse au sein de l'équipe EXPRESSION du Laboratoire IRISA, l'Université de Bretagne-Sud. Je tiens à leur exprimer ma gratitude pour leurs qualités pédagogiques et scientifiques, leur franchise, leur sympathie, leur confiance. J'ai appris beaucoup auprès d'eux. Je leur suis également reconnaissant pour leur écoute, leur partage et leur soutien dans les moments difficiles. J'ai pris un grand plaisir à travailler sous leur direction.

Je voudrais aussi remercier les rapporteurs de cette thèse : Madame Sophie ROSSET, Directrice de Recherche du Laboratoire LIMSI, CNRS et Monsieur le Professeur Emmanuel MORIN au Laboratoire d'Informatique de Nantes-Atlantique, l'Université de Nantes pour l'intérêt qu'ils ont porté à mon travail.

Mes remerciements s'adressent également à Monsieur Gwénolé LECORVÉ de l'Université de Rennes 1 pour avoir accepté d'examiner mon travail et de participer au jury.

Je souhaite remercier tous les membres du laboratoire IRISA, Lab-STICC, ENSIBS : les enseignants, techniciens, administratifs et doctorants qui m'ont aidé et accompagné dans mon travail durant ces quatre années en France.

Je n'oublie pas non plus tous les amis de France qui nous ont aidés, ma famille et moi : Brigitte ENQUEHARD, Evelyne BOUDOU, Alain BOUDOU, Lucien MOREL, Gildas TRÉGUIER, Sylvain CAILLIBOT..., les étudiants vietnamiens et les familles vietnamiennes de Lorient.

Pour terminer, je remercie du fond du cœur mes beaux-parents NONG Quoc Chinh - TRAN Thi Doan, mes parents VU The Huan - LE Thi Nhi et tous les membres de ma famille qui m'ont toujours soutenu, tout au long de ma vie, de mes études, sans lesquels je n'en serais pas là aujourd'hui. Ma reconnaissance va surtout à mon épouse NONG Thi Quynh Tram et à nos deux enfants VU Quynh Maï et VU Haï Minh qui sont toujours à mes côtés et me donnent la force de relever les défis.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Actuellement très présente dans de nombreux domaines du Traitement Automatique des Langues (TAL), l'utilisation de modèles vectoriels statistiques pour étudier la sémantique repose sur l'hypothèse de la *sémantique statistique*, selon laquelle *"the statistical patterns of human word usage can be used to figure out what people mean, at least to a level sufficient for information access"* (les modèles statistiques de l'usage qui est fait des mots peuvent être utilisés pour comprendre ce que les gens disent, tout au moins suffisamment pour accéder à l'information) [1].

Les travaux qui ont été menés au cours de ce doctorat avaient pour objectif initial la réalisation d'un système de résumé automatique concernant un sujet donné à partir d'un ensemble de textes en langue française. Cet objectif a été effectivement atteint comme le montrent les expérimentations décrites à la fin de ce document (cf. page 69) ; cependant, l'essentiel des travaux a été consacré à la conception d'un sous-module du système consacré à l'évaluation de la similarité entre phrases. En l'occurrence, il s'est agi de mesurer jusqu'à quel point ces phrases « parlent de la même chose » et relatent les mêmes faits ou actes. Nous avons choisi de référencer ce sous-module sous l'appellation WikiRI en référence aux modèles et ressources, introduites ci-après, sur lesquels il est fondé.

La tâche qui consiste à mesurer la similarité entre deux phrases ou textes courts (STS : Semantic Textual Similarity) est utilisée, avec des acceptions du mot même de similarité qui peuvent varier sensiblement, dans plusieurs domaines importants du Traitement Automatique des Langues (TAL), au nombre desquels on peut citer la recherche d'informations (Balasubramanian et al. [2007]), la catégorisation de

1. Cité par Turney *et al.* (2010)

textes (Ko et al. [2002]), le résumé de texte (Erkan and Radev [2004]), la traduction automatique, etc.

Comparer les mots ou n-grammes communs entre deux textes constitue une première approche pour mesurer leur similarité (Hirao et al. [2005], Lin [2004]). Cependant, elle ne tient compte, ni des relations sémantiques entre les mots ou groupes de mots d'un même texte, ni de la similarité sémantique entre mots distincts des deux textes (synonymie, paraphrase, etc.). Pour pallier ce manque, le TAL peut s'appuyer sur l'hypothèse distributionnelle avancée par des linguistes tels que Harris [1954] et Firth [1957] selon laquelle *les mots qui apparaissent dans des contextes similaires ont potentiellement des significations similaires* et *"You shall know a word by the company it keeps"* (on peut connaître un mot à partir de ses fréquentations). Ainsi, beaucoup d'approches, comme par exemple LSA (Deerwester et al. [1990]), s'appuient sur l'étude statistique de gros corpus de la langue. En analyse distributionnelle, le modèle initial consiste à construire des matrices *termes*×*contextes* dont les éléments sont une mesure de co-occurrence. Les détails de ces représentations sont décrits dans le chapitre 2.

Le système global de résumé automatique de textes que nous voulons construire doit être robuste, générique et aisément portable et il doit être utilisable pour la langue française. Le choix a donc été fait de faire reposer le système WikiRI sur le modèle vectoriel du Generalized Vector Space Model (GVSM) (Wong et al. [1985]) et d'utiliser l'encyclopédie Wikipédia comme ressource linguistique (cf. page 30). Ainsi, les termes y sont représentés comme des vecteurs dans la base des concepts définis à partir des articles de Wikipédia. Pour remédier aux problèmes posés par le nombre d'articles présents dans cette encyclopédie et sa constante augmentation, nous proposons une représentation vectorielle de la sémantique des termes qui utilise une approche basée Random Indexing (RI) (cf. page 32). Le chapitre 2 explique ces choix parmi ceux que l'on peut trouver actuellement dans l'état de l'art. Les chapitres suivants décrivent nos expérimentations.

Un argument souvent avancé est que l'utilisation d'une encyclopédie telle que Wikipédia introduit du « bruit » dans la sémantique des termes, dans la mesure où chacun d'entre eux se voit utilisé dans toutes ses significations possibles (Gottron et al. [2011]). Le chapitre 3 décrit une première expérimentation que nous avons menée pour définir les vecteurs de termes en sélectionnant les articles de Wikipédia en fonction du domaine concerné, une solution qui n'a pas donné les résultats escomptés. Dans la suite, la totalité de Wikipédia a été utilisée.

Pour le calcul de la similarité entre phrases, deux versions de WikiRI ont été implémentées : la première (WikiRI_1) calcule un vecteur sémantique de phrase à partir des vecteurs sémantiques des termes qui la composent (Chatterjee and Mohan [2007]). Dans cette approche classique, des modifications sont proposées dans les calculs des vecteurs de termes pour corriger le « bruit » engendré par l'utilisation d'une ressource linguistique aussi encyclopédique que Wikipédia : elles sont détaillées dans la section 4.1. La seconde version (WikiRI_2) calcule la similarité entre deux énoncés en comparant directement les similarités entre les termes qui les composent, suivant une méthode proche de celle utilisée par Mihalcea et al. [2006]. L'introduction d'informations syntaxiques a donné lieu à plusieurs variantes de WikiRI_2. Différentes approches ont en effet été testées pour compléter les informations données par une simple approche « sac de mots ». Elles sont décrites dans le chapitre 4. En revanche, pour ne pas rendre plus coûteuse la mise en œuvre du système, aucune expérimentation n'a été menée concernant la construction de matrices liées aux mots prédicatifs. Ainsi, les expérimentations ne reposent que sur les vecteurs de termes construits à partir de Wikipédia et des informations syntaxiques données par des POS Taggers ou des parseurs.

Le chapitre 5 décrit les résultats du système et de ses variantes. Les évaluations concernant la similarité entre phrases ont été effectuées sur des ensembles de données en français que nous avons construites spécifiquement pour évaluer notre approche et sur les données de SemEval 2014 pour l'anglais.

Enfin le chapitre 6 est consacré aux expérimentations de résumé automatique multi-documents que nous avons pu mener en utilisant les résultats de similarité rendus par le système WikiRI. Le système obtient des résultats globalement satisfaisants mais surtout, les comparaisons entre WikiRI_1, WikiRI_2 et ses variantes donnent des indications concernant les pistes d'amélioration et aussi les limites de l'approche.

Chapitre 2

Représentation sémantique d'un terme

La sémantique est une branche de la linguistique qui étudie les signifiés, c'est-à-dire ce dont parle un énoncé. Elle recouvre notamment l'étude du sens lexical (ou sens des mots) et l'étude du sens de combinaisons de mots, de phrases ou de textes.

On distingue souvent la sémantique de la syntaxe qui est la branche de la linguistique qui étudie les signifiants, c'est-à-dire la façon dont les mots se combinent pour former des phrases ou des énoncés.

2.1 Quelques approches de la sémantique lexicale

Au contraire de la syntaxe, la sémantique lexicale se caractérise par une grande diversité d'approches. Certaines se focalisent sur la représentation du sens, comme celles basées sur les graphes ou les espaces vectoriels tandis que d'autres s'attachent, en plus, à rendre compte du processus de calcul ou de constitution du sens, comme celles basées sur la logique mathématique ou les modèles neuronaux. Nous présentons ci-dessous quelques unes de ces approches.

2.1.1 Modèles graphiques

Les réseaux sémantiques (Collins and Quillian [1969]) sont des graphes orientés, formés de noeuds (figurant des concepts qui peuvent être des mots, des groupes de mots, des catégories...) reliés par des arcs qui expriment les relations sémantiques entre concepts (par exemple, un chat est un mammifère, un chat a des oreilles).

Le sens d'un concept est induit par le nombre et le type de connexions qu'il entretient avec ses voisins. Dans ce cadre, la similarité entre deux concepts est fonction des longueurs de chemin qui séparent leurs noeuds et on s'attend à ce que des concepts sémantiquement proches soient reliés par des chemins plus courts.
Plusieurs systèmes (WordNet Fellbaum [1998], FrameNet Ruppenhofer et al. [2006], ...) utilisent le graphe comme paradigme de représentation lexicale. L'exemple le plus connu est le thésaurus WordNet qui propose une organisation hiérarchique du sens lexical. Chaque noeud y correspond à une liste de synonymes (synset pour synonym set) qui définissent une acception ou un usage particulier d'un mot. Par exemple, le nom commun "boy" admet trois synsets différents qui correspondent aux trois acceptions : jeune de sexe masculin, fils et référence informelle à un adulte. Le lexique de WordNet est découpé en quatre grandes catégories lexicales (noms, verbes, adjectifs, adverbes) et une variété de relations sémantiques permettent d'organiser le sens des mots, dont l'hyperonymie (animal est un hyperonyme de chat et chat est un hyponyme d'animal), la méronymie (qui lie une partie de l'objet au tout), l'antonymie (qui lie des opposés), etc. La qualité de représentation des mots est très variable selon les catégories ; l'architecture hiérarchique de Wordnet est en effet plus adaptée à la représentation des noms qu'à celles des verbes, des adjectifs ou des adverbes.
Du point de vue de l'implémentation, la construction des réseaux sémantiques est peu automatisée. Elle se fait essentiellement "à la main" par des annotateurs qui décident a priori quelles relations sont plus pertinentes pour représenter le sens. La base lexicale Wordnet par exemple comporte plus de 115 000 synsets annotés manuellement et on réalise aisément l'ampleur des efforts nécessaires pour sa réalisation ou sa maintenance. Des travaux plus récents (Steyvers [2005]) créent des réseaux sémantiques à partir de normes d'association de mots (Nelson et al. [2004]) ; le vocabulaire couvert reste cependant limité.
En plus de la quantité de travail nécessaire à leur création et leur adaptation, les représentations discrètes de type de Wordnet posent d'autres problèmes comme la

difficulté voire l'impossibilité de leur mise à jour, leur subjectivité due à l'annotation manuelle et leur peu de disponibilité dans d'autres langues que l'anglais.

2.1.2 Modèles d'espaces vectoriels et modèles neuronaux

Il existe de nombreux modèles vectoriels dont LSA (Latent Semantic Analysis) (Landauer et al. [1998]), HAL (Hyperspace Analogue to Language) (Burgess and Lund [1997]) pour les plus célèbres, et COALS (Rohde et al. [2006]) et Hellinger-PCA (Lebret and Collobert [2014]) pour les plus récents. Ces modèles ne considèrent pas une organisation a priori des concepts ou de la sémantique des unités lexicales (comme dans Wordnet) mais ils établissent les liens sémantiques entre mots à partir de leur emploi dans des corpus de textes qui peuvent atteindre plusieurs millions de mots.

Étant donné un corpus, on construit une matrice de cooccurrences qui croise les mots du vocabulaire (en ligne) avec les concepts (en colonne). Ceux-ci peuvent être des mots (comme dans HAL), des phrases, paragraphes, documents entiers (comme dans LSA) ou encore des fenêtres de mots autour du mot cible. On parle dans ce dernier cas de concept local. La matrice de cooccurrences obtenue présente le double désavantage d'être creuse et de grande taille et sa factorisation, souvent par décomposition en valeurs singulières (SVD), est nécessaire afin de donner une représentation dense et en faible dimension des mots.
L'évaluation de la similarité sémantique s'appuie ici sur l'hypothèse distributionnelle selon laquelle des mots qui ont des contextes similaires sont sémantiquement proches. Tels qu'ils sont construits, les vecteurs repésentatifs des mots rendent effectivement compte de leurs contextes d'utilisation et la similarité entre mots est donc évaluée à partir de la similarité entre vecteurs (souvent le cosinus de l'angle qu'ils forment).
Enfin, l'utilisation de contextes globaux permet une représentation des mots par thématique générale (des mots thématiquement proches ont des vecteurs représentatifs proches) alors que des contextes plus locaux permettent de capturer à la fois de l'information syntaxique (POS - Part Of Speech) et de l'information sémantique.

Les modèles vectoriels sont tout à fait adaptés à une implémentation informatique et leur mise en oeuvre ne nécessite qu'un corpus de taille suffisante et des algorithmes matriciels classiques. Le coût de ces algorithmes devient cependant prohibitif pour de très gros corpus ; le coût de la SVD, par exemple, est en $O(mn^2)$ pour des matrices $m \times n$ avec $n < m$. Par ailleurs, une fois le corpus traité, il est difficile d'y incorporer de nouveaux mots ou documents.

Plutôt que de construire des vecteurs creux de grande dimension et de les projeter dans des espaces vectoriels de dimension réduite, d'autres modèles s'attachent à représenter directement les mots par des vecteurs de faible dimension, comme la méthode du Random Indexing et les modèles prédictifs à base de réseaux de neurones.
La méthode du Random Indexing telle qu'introduite par Sahlgren [2005a] procède en deux temps. Les concepts sont d'abord représentés par des vecteurs aléatoires de faible dimension puis les vecteurs représentatifs des mots sont calculés par sommation des vecteurs des concepts auxquels ils sont associés. Nous utilisons dans la suite de nos travaux une variante pondérée du Random Indexing qui utilise Wikipédia comme ressource linguistique.
De leur côté, les modèles prédictifs utilisent des réseaux de neurones pour apprendre les représentations vectorielles des mots à partir de corpus d'apprentissage. Les vecteurs induits sont denses, de faible dimension et chaque direction représente une caractéristique latente du mot, sensée capturer des propriétés syntaxiques et sémantiques. On parle de représentations distribuées. Le modèle de Rumelhart et al. [1986] apprenait déjà à représenter des mots par rétro-propagation des erreurs mais le véritable essor des modèles neuronaux a démarré avec Bengio et al. [2003]. Une limitation des premiers modèles neuronaux était la dépendance linéaire des temps d'exécution à la taille du vocabulaire, dans les étapes d'apprentissage et de test mais les nouvelles approches (Morin and Bengio [2005] ; Collobert and Weston [2008] ; Mnih and Hinton [2008]) qui utilisent des réseaux de neurones à l'architecture plus complexe, ont permis de passer à de gros corpus d'apprentissage.

Un modèle plus simple et plus rapide, implémenté dans l'outil word2vec, a été récemment introduit par Mikolov et al. [2013b]. Il utilise deux modèles prédictifs basés sur des réseaux de neurones à simple couche : skip-gram et Continuous Bag Of Words (CBOW). Étant donnée une fenêtre de n mots autour d'un mot w, le modèle skip-gram prédit ses mots voisins dans la fenêtre fixée. Le modèle CBOW permet ensuite de prédire le mot cible w, étant donnés ses voisins dans la fenêtre.

Des modèles similaires ont été proposés par Mnih and Kavukcuoglu [2013] et Levy and Goldberg [2014]. En plus de leur simplicité et de leur rapidité d'exécution, ces modèles présentent l'avantage d'incorporer aisément de nouvelles phrases ou documents dans le corpus et d'ajouter de nouveaux mots au vocabulaire. Cependant, le fait de ne considérer les contextes qu'au travers de petites fenêtres de mots limite l'accès à l'information portée par la répétition des données.

Un modèle récent qui réunit les deux approches factorisation de matrice et modèles prédictifs, a été récemment proposé par Pennington et al. [2014]. GloVe (pour Global Vectors) est un modèle d'apprentissage non supervisé qui exploite l'ensemble de l'information portée par le corpus et non plus la seule information portée par une fenêtre de mots. Les entrées non nulles de la matrice globale de cooccurrences mot-mot sont utilisées pour entrainer un modèle log-bilinéaire qui calcule les vecteurs représentatifs des mots. GloVe s'est avéré rapide et efficace dans diverses tâches comme la recherche de similarité, la reconnaissance d'entités nommés ou encore l'analyse de sentiments, y compris avec de petits corpus.

2.1.3 Modèles géométriques

La modélisation géométrique est une modélisation continue qui associe à un mot non plus un atome ou plusieurs atomes de sens (vecteur ou noeuds d'un graphe) mais un domaine dans un espace multi-dimensionnel.
Dans le modèle des Atlas sémantiques (Ploux [1997] ; Ploux and Victorri [1998]), les entrées de plusieurs dictionnaires sont organisées en cliques de similarité, c'est-à-dire en ensembles maximaux de mots tous synonymes les uns des autres. D'autres types de cliques peuvent être envisagés comme les cliques de contexte (Ji et al. [2003]).
D'un point de vue mathématique, une clique est un sous-graphe maximal complet connexe et il n'existe aucun autre mot dans le lexique qui puisse la subdiviser.

Une clique figure un sens assez précis de contexte et d'emploi, ainsi qu'illustré par les cliques qui partagent le mot vedette *faible* :

— abattu, anéanti, faible, fatigué, las
— abattu, déprimé, faible, fatigué
— affaibli, anémique, anémié, débile, faible
— amour, faible, goût, inclination, passion, penchant

— faible, faiblesse, goût, inclination, préférence
— dérisoire, faible, insignifiant, minime, négligeable, petit

Les cliques organisent le sens en valeurs type (physique, émotionnelle, perceptuelle...) et comme chacune est connectée à la suivante par un synonyme commun, il est possible d'envisager une transition continue du sens à des niveaux sémantiques subtils.

Étant donné un corpus, une liste de cliques est établie puis une matrice de présence/absence croise les mots en colonne et les cliques en ligne. Une analyse factorielle des correspondances (Benzécri [1980]) permet ensuite de représenter les cliques par des vecteurs denses et de faible dimension. À chaque clique est affecté un point d'un espace affine multidimensionnel et chaque mot du corpus est représenté par l'enveloppe des points associés aux cliques qui le contiennent ; ainsi, à chaque mot est associé une zone de l'espace de représentation. Enfin, un algorithme de classification permet de distinguer à partir du nuage de points formé par les cliques les différentes valeurs du mot. La distance entre mots est mesurée par la mesure du χ_2 qui est particulìrement adaptée pour rendre compte de leur organisation géométrique.

2.1.4 Modèles logico-algébriques

La grammaire générative est un modèle développé par Chomsky [1957] pour théoriser l'aptitude humaine à produire des phrases grammaticales. Pour Chomsky, sémantique et syntaxe sont indépendantes et il décrit la syntaxe comme un système formé d'un vocabulaire, d'axiomes et de règles d'inférence.

À l'inverse, la grammaire de Montague [1974] repose sur l'existence d'un homomorphisme entre la syntaxe et la sémantique, en tant qu'algèbres : à chaque règle syntaxique est associée une règle sémantique. Elle suppose également que le sens d'un énoncé peut s'obtenir en composant celui de ses constituants. Montague propose ainsi des représentations sémantiques pour une partie de la langue anglaise en utilisant des formules de la logique des prédicats du premier ordre.

Plus récemment, Pustejovsky [1998] a développé le modèle du Lexique génératif (LG) pour répondre à la critique du traitement de la polysémie dans les approches classiques. Au contraire des modèles énumératifs qui tâchent de répertorier tous les sens possibles des mots (comme dans Wordnet), le Lexique génératif est un modèle

explicatif de la polysémie, qui articule sémantique et syntaxe pour déterminer le sens en contexte au moyen d'un ensemble d'axiomes et des règles de dérivation. J. Pustejovsky a choisi le lambda-calcul pour réaliser son projet.

2.2 Les espaces vectoriels sémantiques

Une propriété intéressante des modèles d'espaces vectoriels (ou VSM pour Vector Space Models) est qu'ils extraient automatiquement la connaissance à partir de corpus et qu'ils nécessitent moins de travail que d'autres approches de la sémantique qui s'appuient sur des bases de connaissances annotées manuellement ou des ontologies. Par exemple, la ressource principale utilisée par le système VSM de Rapp [2003] pour la mesure de similarité entre mots est le British National Corpus (BNC) tandis que des systèmes non VSM comme ceux de Hirst and St-Onge [1998] ; Leacock and Chodrow [1998] ; Jarmasz and Szpakowicz [2003] utilisent des lexiques comme WordNet ou Roget's Thesaurus.
Les VSM ont largement démontré leur efficacité sur des tâches de mesure de similarité entre mots, phrases et documents ; la plupart des moteurs de recherche les utilisent d'ailleurs pour mesurer la similarité entre une requête et un document (Manning et al., 2008). On les retrouve également comme modèles de représentation dans les algorithmes leaders de mesure du lien sémantique (Pantel and Lin [2002a] ; Rapp [2003] ; Turney et al. [2003]) et des algorithmes de mesure de similarité sémantique (Lin and Pantel [2001] ; Turney [2006] ; Nakov and Hearst [2008]).

Un autre intérêt des VSM est leur relation avec l'hypothèse distributionnelle qui postule que les mots qui occurrent dans des contextes similaires tendent à avoir des sens similaires (Wittgenstein [1953] ; Harris [1954] ; Weaver [1955] ; Firth [1957] ; Deerwester et al. [1990]). Cette hypothèse trouve une formalisation mathématique avec la représentation des mots par des vecteurs, matrices ou tenseurs d'ordre supérieur.

L'utilisation de vecteurs et de matrices ne définit cependant pas à elle-seule la notion de VSM. En effet, les composantes des vecteurs doivent dériver de fréquences d'événements, comme le nombre de fois où un mot apparait dans un contexte donné. Ainsi, un lexique ou une base de connaissance peuvent être vus comme des graphes et un graphe peut être représenté par une matrice d'adjacence mais ceci

n'implique pas qu'un lexique est un VSM, parce qu'en général, les éléments de la matrice d'adjacence ne dérivent pas de fréquences d'événements. Le dénominateur commun que sont les fréquences d'événements apporte de l'unité à la variété de VSM et les connecte explicitement à l'hypothèse distributionnelle ; de plus, il évite la trivialité en excluant de nombreuses représentations matricielles possibles.

2.2.1 Différentes représentations sémantiques

L'idée qui unifie les VSM est l'hypothèse sémantique statistique : « Les modèles statistiques de l'usage d'un mot peuvent être utilisés pour comprendre leur sens ». Suivant Turney and Pantel [2010], cette hypothèse générale se décline en plusieurs autres hypothèses spécifiques : l'hypothèse du sac-de-mots, l'hypothèse distributionnelle, l'hypothèse distributionnelle étendue, et l'hypothèse de relation latente.

2.2.1.1 Matrice terme-document et similarité entre documents

On désigne par sac (bag) un ensemble de mots dans lequel les répétitions sont autorisées et où l'ordre des mots ne compte pas. Par exemple, $\{a, a, b, c, c, c\}$ est un sac contenant les mots a, b, c et les sacs $\{a; a; b; c; c; c\}$ et $\{c; a; c; b; a; c\}$ sont équivalents. On peut représenter le sac $\{a; a; b; c; c; c\}$ par le vecteur $x = (2; 1; 3)$ en posant que le premier élément de x est la fréquence de a dans le sac, le second élément est la fréquence de b et le troisième est la fréquence de c.
Un ensemble de sacs peut être représenté par une matrice dans laquelle chaque colonne correspond à un sac, chaque ligne correspond à un unique mot du vocabulaire et le terme au croisement de la ligne i et de la colonne j correspond à la fréquence du i-ème mot dans le j-ième sac.
Dans la suite, on dispose d'une collection de documents (textes, pages web...) et chaque document est considéré comme un sac-de-mots. La matrice qui croise les documents et les termes du vocabulaire est appelée matrice terme-document.
L'hypothèse du sac-de-mots a été introduite par Salton et al. [1975] et elle est à l'origine de l'utilisation des VSM en recherche d'information. Cette hypothèse avance qu'on peut estimer la pertinence de documents par rapport à une requête en représentant les documents et la requête comme des sacs-de-mots. L'idée sous-jacente est que les fréquences des mots dans un document rendent compte jusqu'à un certain point du sens du document, de ce à quoi il se rapporte ; il apparait

alors légitime de s'en servir pour évaluer la pertinence du document par rapport à une requête. Une justification intuitive de cette approche est que le sujet du document influence très vraisemblablement son auteur dans le choix des mots qu'il emploie. Si deux documents portent sur des sujets similaires, alors leurs vecteurs représentatifs dans la matrice terme-document, auront tendance à avoir les mêmes distributions de fréquences.

Supposons que notre collection contienne n documents et m termes uniques. La matrice terme-document X correspondante est alors constituée de m lignes (une ligne par unique terme du vocabulaire) et n colonnes (une colonne pour chaque document). Soit w_i le i-ème terme du vocabulaire et d_j le j-ème document de la collection. Le vecteur ligne $x_{i\cdot}$ contient n éléments, un pour chaque document et le vecteur colonne $x_{\cdot j}$ contient m éléments, un pour chaque terme du vocabulaire. Si X est une simple matrice de fréquences, l'élément x_{ij} de X est la fréquence du terme w_i dans le document d_j. En général, la matrice X est creuse (la plupart de ses éléments sont nuls) du fait que la plupart des documents n'utilisent qu'une petite fraction de l'ensemble du vocabulaire. Si on choisit au hasard un terme w_i et un document d_j, il est probable que le mot w_i n'occurre nulle part dans le document d_j, et ainsi $x_{ij} = 0$. La distribution des fréquences dans $x_{i\cdot}$ est une signature du i-ème terme w_i ; de même la distribution des fréquences dans $x_{\cdot j}$ est une signature du document d_j. Ces modèles nous indiquent donc ce à quoi se rapportent le mot ou le document.

Le vecteur $x_{\cdot j}$ est une représentation assez grossière du document d_j. En effet, il nous indique à quelles fréquences y apparaissent les mots, mais leur ordre séquentiel est perdu et la structure des expressions, phrases, paragraphes est perdue. Cependant, ces vecteurs semblent capturer un aspect important de la sémantique et malgré la grossièreté de l'approche, les moteurs de recherche qui en sont de grands utilisateurs, donnent de bons résultats.

2.2.1.2 Matrice mot-contexte et similarité entre mots

Salton et al. [1975] se sont limités dans leurs travaux à la mesure de similarité entre documents. Ils considèrent la requête faite au moteur de recherche comme un pseudo-document et la pertinence d'un document par rapport à cette requête est évaluée par la similarité de leurs vecteurs représentatifs. Deerwester et al. [1990] ont étendu leur approche à la mesure de similarité entre mots en se focalisant, non

plus sur les vecteurs colonne de la matrice terme-document mais sur les vecteurs ligne.

Le document n'est toutefois pas toujours la taille optimale de texte pour mesurer la similarité entre mots et d'autres contextes peuvent s'avérer plus pertinents comme des mots, des expressions, des phrases, des paragraphes ou d'autres contextes moins communs comme des séquences de caractères ou des patrons. La thèse de Sahlgren [2006] offre une bonne revue de la diversité des contextes utilisés dans la littérature, dont des fenêtres de mots autour du mot cible (Lund and Burgess [1996]), des dépendances grammaticales (Lin [1998] ; Padò and Lapata [2007]), etc.

L'utilisation des VSM pour la mesure de similarité entre mots s'appuie sur l'hypothèse distributionnelle qui affirme que des mots qui occurrent dans des contextes similaires tendent à avoir des sens similaires (Harris [1954]). Ainsi, des vecteurs ligne similaires dans la matrice mots-contextes indiquent des sens de mots similaires. L'idée sous-jacente à cette hypothèse est que le sens recherché du mot est son usage en contexte plutôt que sa signification littérale.

L'idée que l'usage que l'on fait d'un mot peut révéler sa sémantique est ancienne. Elle était déjà implicite dans certains des travaux de Wittgenstein [1953] : "Don't look for the meaning, but for the use" et Harris [1954] considérait le sens comme une fonction de distribution. Dans ses travaux en traduction automatique, Weaver [1955] proposait de désambiguïser les mots en se basant sur la fréquence des mots-contexte proches et Firth [1957] affirmait pour sa part : "You shall know a word by the company it keeps." Les travaux de Deerwester et al. [1990] ont finalement permis de mettre en œuvre dans un algorithme pratique les intuitions de Wittgenstein [1953], Harris [1954], Weaver [1955] et Firth [1957].

2.2.1.3 Matrice paire-patron et similarité relationnelle

Dans une matrice paire-patron, les vecteurs ligne correspondent à des paires de mots, comme (maçon-pierre) et (charpentier-bois) et les vecteurs colonne aux patrons où les paires de mots co-occurrent, comme "X coupe Y" ou "X travaille avec Y". Lin and Pantel [2001] ont introduit les matrices paire-patron dans le but de mesurer la similarité sémantique entre patrons, dans une tâche de détection de paraphrases. Leur algorithme permet par exemple, d'identifier le patron "Y est résolu par X" comme similaire au patron "X résoud Y". Lin and Pantel [2001] ont étendu la notion d'hypothèse distributionnelle aux patrons : "les patrons qui co-occurrent

avec des paires de mots similaires ont tendance à avoir des sens similaires". Ainsi, les patrons "X résoud Y" et "Y est résolu par X" tendent à co-occurrer avec des paires $X - Y$ similaires ; ce qui suggère que ces patrons ont des sens similaires.
Turney et al. [2003] ont quant à eux utilisé la matrice paire-patron pour évaluer la similarité sémantique de relations entre paires de mots ; c'est-à-dire la similarité de vecteurs ligne. Par exemple, les paires (maçon-pierre), (charpentier-bois), (potier-argile), (souffleur de verre-verre) partagent la même relation sémantique "artisan-matériel". Dans chaque cas, le premier membre de la paire est un artisan qui fabrique des objets à partir du matériel spécifié dans le second membre de la paire. Les paires tendent à co-occurrer dans des patrons similaires, comme "X a utilisé Y pour" et "X a façonné Y en".
L'hypothèse de relation latente consiste à dire que les paires de mots qui co-occurrent dans des patrons similaires tendent à avoir des relations sémantiques similaires (Turney [2008]) ; soit que les paires de mots qui ont des vecteurs ligne similaires dans une matrice paire-patron tendent à avoir des relations sémantiques similaires. C'est l'inverse de l'hypothèse distributionnelle étendue qui dit que les patrons avec des vecteurs colonnes similaires dans la matrice paire-patron tendent à avoir des sens similaires.

2.2.1.4 Autres représentations

Les matrices terme-document, mot-contexte, paire-patron ne sont pas exhaustives des VSM. On peut considérer des matrices triplet-patron, pour mesurer la similarité sémantique entre des triplets de mots. Alors que la matrice paire-patron doit avoir une ligne (maçon,pierre) et une colonne "X travaille avec Y", une matrice triplet-patron pourrait avoir une ligne (maçon, pierre, maçonnerie) et une colonne "X utilise Y pour construire Z". Cependant, les n-uplets de mots deviennent très rares lorsque n augmente. Par exemple, des expressions qui contiennent à la fois maçon, pierre et maçonnerie sont moins fréquentes que des expressions qui ne contiennent que maçon et pierre. Une matrice triplet-patron est donc bien plus creuse qu'une matrice paire-patron. La quantité de texte nécessaire pour rendre ces matrices utiles augmente rapidement lorsque n augmente ; aussi, il est préférable de scinder les n-uplets en paires. Turney [2008] décompose les triplets (a,b,c) en paires (a,b), (a,c) et (b,c) et la similarité entre les deux triplets (a,b,c) et (d,e,f) est estimée à partir des similarités des paires induites. Une matrice paire-patron relativement dense peut donc servir de substitut à une matrice triplet-patron creuse. On

peut également aller au delà des matrices en considérant des tenseurs qui sont des généralisations de matrices (Kolda and Bader [2009] ; Acar and Yener [2009]) : un scalaire est un tenseur d'ordre 0, un vecteur est un tenseur d'ordre 1 et une matrice est un tenseur d'ordre 2. Un tenseur d'ordre 3 est appelé tenseur d'ordre supérieur. Chew et al. [2007] utilisent un tenseur d'ordre 3 terme-document-langage pour de la recherche d'information multilingue, Turney [2007] utilise un tenseur mot-mot-patron pour mesurer la similarité entre mots et Van de Cruys [2009] utilise un tenseur verbe-sujet-objet pour apprendre les classes d'arguments préférentiels de verbes. Dans le tenseur de Turney [2007], les lignes correspondent aux mots dans le questionnaire à choix multiple de synonymes du TOEFL, les colonnes correspondent aux mots du Basic English (Ogden [1930]) et la troisième dimension correspond aux patrons qui relient les lignes et colonnes ; on a ainsi un tenseur d'ordre 3 mot-mot-patron. Un mot du questionnaire TOEFL est représenté par une matrice mot-patrons dont les éléments correspondent à l'ensemble des patrons qui relient le mot considéré à tout autre mot du Basic English, et la similarité de deux mots TOEFL est calculée par comparaison des deux matrices qui leur sont associées.

2.2.2 VSM et types de similarité

Les travaux cités plus haut utilisent les notions de lien sémantique, de similarité taxonomique et d'association sémantique.

— La notion de lien sémantique (Budanitsky and Hirst [2001]) correspond à la similarité attributionnelle en sciences cognitives (Gentner [1983]). Deux mots sont sémantiquement liés s'il existe une relation sémantique quelconque entre eux, comme une relation de synonymie, de méronymie (*voiture* et *volant*), d'antonymie (*chaud* et *froid*), un lien fonctionnel entre mots ou encore si ces mots sont fréquemment associés (*papier* et *crayon*). Par exemple, les antonymes *chaud* et *froid* ont un grand degré de similarité attributionnelle puisqu'ils se rapportent tous deux à des températures, noir et blanc à des couleurs, etc.

— La notion de similarité taxonomique concerne des mots qui partagent un hyperonyme (*voiture* et *vélo* sont taxonomiquement similaires car ils partagent l'hyperonyme *véhicule*)(Resnik [1995]). La similarité taxonomique est un cas particulier de lien sémantique.

— Des mots sont sémantiquement associés s'ils tendent à co-occurrer fréquemment, comme *abeille* et *miel* (Chiarello et al. [1990]). Des mots peuvent être taxonomiquement similaires et sémantiquement associés (*médecin* et *infirmière*), taxonomiquement similaires mais non associés sémantiquement (*cheval* et *platypus*), sémantiquement associés mais non taxonomiquement similaires (*berceau* et *bébé*), et ni sémantiquement associés ni taxonomiquement similaires (*calcul* et *bonbon*).

Dans les modélisations VSM, les matrices mot-contexte sont adaptées à la mesure du lien sémantique entre mots (similarité attributionnelle) tandis que les matrices paire-patron sont adaptées à la mesure de la similarité relationnelle entre paires de mots. Par exemple, les mots *chien* et *loup* ont une similarité attributionnelle élevée et les paires de mots *(chien,aboie)* et *(chat,miaule)* ont une similarité relationnelle élevée (Turney [2006]). Notons encore que la similarité relationnelle ne peut être réduite ou inférée à partir de la similarité attributionnelle. Par exemple, *maçon*, *charpentier*, *potier* et *souffleur de verre* sont des mots similaires (ce sont tous des artisans), comme le sont les mots *argile*, *pierre* et *verre* (ce sont tous des matériaux utilisés par des artisans), mais on ne peut en inférer que *(maçon,verre)* et *(charpentier,argile)* aient des relations similaires.

2.3 Traitements mathématiques des VSM

Le processus de construction d'un VSM se compose de quatre étapes : 1) après pré-traitement du corpus, tokenisation et de manière non systématique lemmatisation des termes et/ou de leurs contextes, filtrage des mots outils (déterminants, pronoms, etc.), une matrice de fréquences est générée. 2) Les fréquences brutes relevées sont ensuite ajustées. Les raisons de cette transformation sont diverses et la première est de réduire l'importance des mots communs qui ont de plus grandes fréquences d'occurrence alors qu'ils sont moins informatifs que les mots rares. 3) Lissage de la matrice de manière à réduire le bruit et à la rendre plus dense. 4) Calcul de similarités entre vecteurs. Nous disposons pour cela de nombreuses possibilités en fonction de la tâche à accomplir. Lowe [2001] donne un bon résumé du traitement mathématique des VSM mots-contexte et Bullinaria and Levy [2007a] analysent l'importance de divers facteurs qui entrent en jeu dans chacune des étapes ci-dessus.

2.3.1 Construction de la matrice des fréquences brutes

Un élément de la matrice des fréquences brutes correspond au nombre de réalisations (fréquence) d'un événement : un item (terme, mot, paire de mots) occurre dans une certaine situation (document, contexte, patron). Construire une matrice de fréquences est une simple question de comptage d'événements qui peut néanmoins devenir compliquée en présence de gros corpus. Une approche typique consiste à scanner séquentiellement le corpus et à enregistrer les événements et leurs fréquences dans une table de hashage, une base de données ou un moteur de recherche. On utilise ensuite la structure de données résultante pour générer la matrice de fréquences.

2.3.2 Pondération des fréquences brutes

En théorie de l'information, un événement rare a un contenu informatif plus important qu'un événement commun (Shannon [1948]). En recherche sémantique, les événements rares partagés par deux vecteurs, sont plus discriminants de la similarité entre vecteurs que les événements communs. Par exemple, dans la mesure de similarité sémantique entre les mots souris et rat, les contextes "disséquer" et "exterminer" sont plus discriminants que les contextes "avoir" et "aimer". On cherchera donc à modifier les fréquences brutes relevées dans le corpus afin de donner plus de poids aux événements rares et moins de poids aux événements plus communs.

TF	IDF	Normalisation
1 si $f_{t,d} > 0$ et 0 sinon	1	Aucune
$f_{t,d}$	$\log(N/n_t)$	$v/\|v\|_2$
$\log(1 + f_{t,d})$	$\log(1 + N/n_t)$	$1/u$
$kf_{t,d}/(k + f_{t,d})$, $k \geq 1$	$\log(1 + \max_t N/n_t)$	$1/charlength^{\alpha}$, $\alpha < 1$
$(k + 1)f_{td}/(f_{td} + k)$, $k \geq 0$	$\log((N - n_t)/n_t)$	$1 - k + k\|d\|/lmd$, $0 \leq k \leq 1$
$k + (1 - k)f_{t,d}/\max_t f_{t,d}$, $k \geq 0$		

TABLE 2.1: Quelques pondérations tf, idf et normalisations

La pondération la plus populaire pour des matrices terme-document est la famille de fonctions de pondération tf-idf (term frequency×inverse document frequency) introduite par Spärck Jones [1972]. Un terme obtient un poids élevé dans un document lorsqu'il y est fréquent (son tf est élevé) et qu'il est rare dans d'autres

documents du corpus (son df est bas et donc son idf est élevé). Salton and Buckley [1988] ont défini une large famille de fonctions poids tf-idf qu'ils ont évaluées sur des tâches de recherche d'information, démontrant que les pondérations tf-idf peuvent apporter des améliorations significatives par rapport à des fréquences brutes. Un autre type de pondération, souvent combinée avec la pondération tf-idf, est la normalisation de la longueur du document (Singhal et al. [1996]). En recherche d'information, si la longueur de document est ignorée, les moteurs de recherche tendent à avoir un biais en faveur des document les plus longs ; la normalisation de la longueur des documents corrige ce biais.
La pondération des fréquences brutes peut également être utilisée pour corriger la corrélation entre termes. Par exemple, les termes *journal* et *journaux* sont corrélés et si on ne souhaite pas les rapporter à un même lemme, on peut réduire leurs poids lorsqu'ils co-occurrent dans un document (Church [1995]).
Nous présentons dans la table 2.1 quelques pondérations TF, IDF et normalisations usuelles. On peut les associer librement mais celles qui sont le plus souvent utilisées ensemble apparaissent de la même couleur. On retrouve ainsi, le tf-idf classique (en bleu) :

$$tf\text{-}idf(t,d) = f_{t,d} \times \log(N/n_t)$$

et la fonction de pondération BM25-Okapi (en orange), très utilisée en recherche d'information pour classer des réponses à une requête :

$$tf\text{-}idf_{BM25}(t,d) = \frac{(k_1+1)f_{t,d}}{f_{t,d} + k_2(1-k_2+l|d|/lmd)} \times \log(N/n_t)$$

avec $f_{t,d}$ la fréquence d'occurrence du terme t dans le document d, N le nombre de documents de la collection, n_t le nombre de documents contenant le terme t, $|d|$ la longueur du document d et lmd la longueur moyenne des documents du corpus.
Toutes les fonctions de pondération TF favorisent les termes fréquents dans un document mais elles atténuent différemment l'effet des fréquences élevées ; les figures 2.1 et 2.2 en donnent une illustration. Les fonctions IDF, de leur côté, favorisent les termes rares à l'échelle du corpus comme le montre la figure 2.3. Enfin, la normalisation pivotale, qui est présentée dans la figure 2.4, favorise les documents qui sont plus courts que la moyenne des documents du corpus.

Une alternative au tf-idf est la Pointwise Mutual Information (PMI) (Church and Hanks [1989] ; Turney [2001]) qui est définie ci-dessous et qui donne de bons résultats aussi bien pour des matrices mot-contexte (Pantel and Lin [2002a]) que

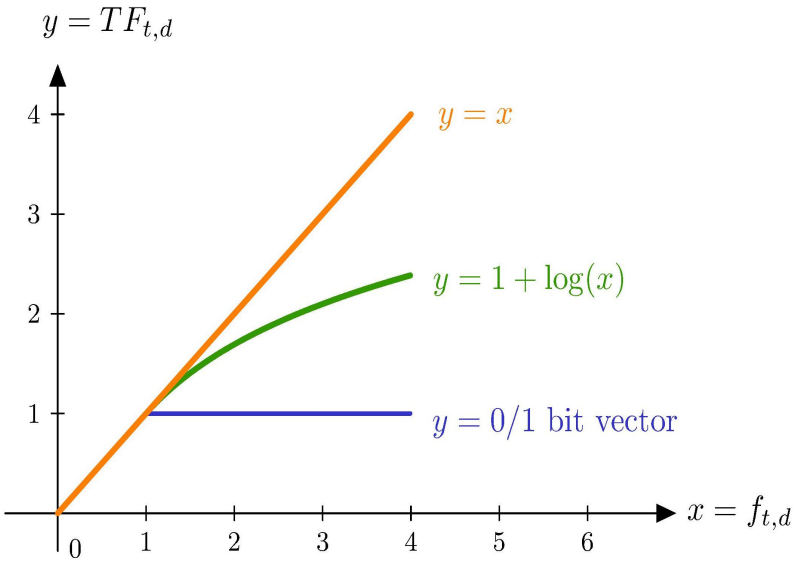


FIGURE 2.1: Pondérations TF

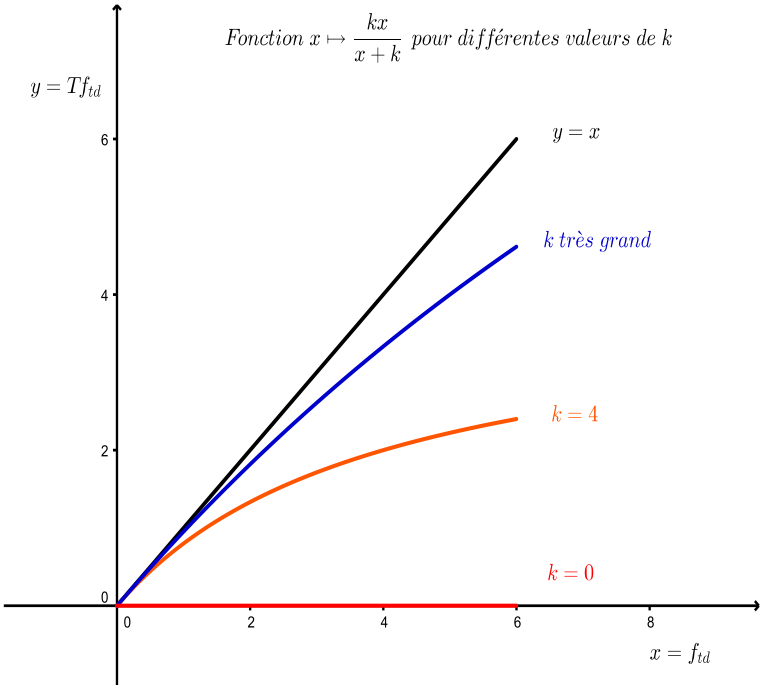


FIGURE 2.2: Pondération $BM25$

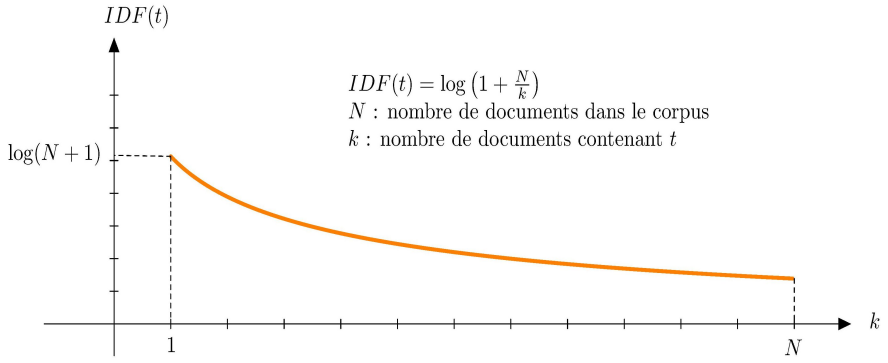


FIGURE 2.3: Pondération IDF

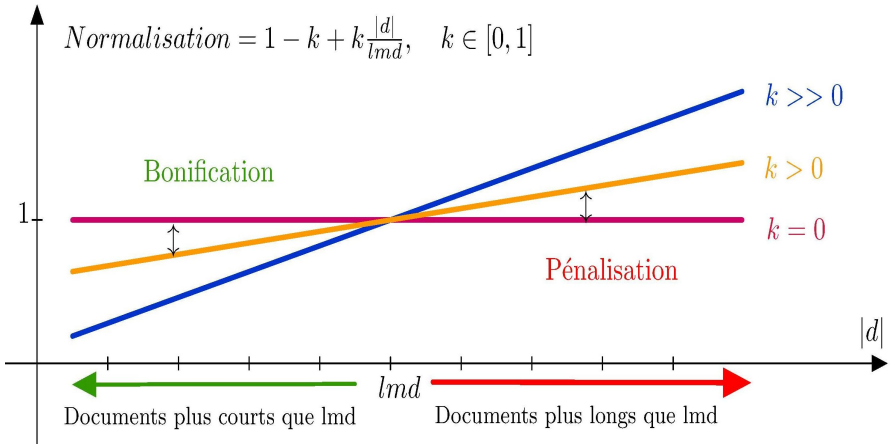


FIGURE 2.4: Normalisation pivot de la longueur des documents

pour des matrices terme-document (Pantel and Lin [2002b]). La positive pointwise mutual information (PPMI) est une variante de la PMI dans laquelle toutes les valeurs négatives sont ramenées à zéro (Niwa and Nitta [1994]). Bullinaria and Levy [2007a] démontre que la PPMI offre des performances meilleures qu'une grande variété de fonctions de pondération pour la mesure de similarité sémantique avec les matrices mot-contexte et Turney [2008] a observé son efficacité avec des matrices paire-patron.

Soit F une matrice de fréquences brutes mot-contexte avec m lignes et n colonnes. La i-ème ligne de F est le vecteur $f_{i\cdot}$ et la j-ième colonne de F est le vecteur colonne $f_{\cdot j}$. La ligne $f_{i\cdot}$ correspond au mot w_i et la colonne $f_{\cdot j}$ correspond au contexte c_j. La valeur de l'élément f_{ij} est le nombre de fois que w_i se trouve dans le contexte c_j. Soit X la matrice qui résulte de l'application de la pondération PPMI à F. La nouvelle matrice X a le même nombre de lignes et de colonnes que la matrice de

fréquences brutes F et la valeur d'un élément x_{ij} de X est définie comme suit :

$$p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{n_r}\sum_{j=1}^{n_c} f_{ij}} \quad (1)$$

$$p_{i*} = \frac{\sum_{j=1}^{n_c} f_{ij}}{\sum_{i=1}^{n_r}\sum_{j=1}^{n_c} f_{ij}} \quad (2)$$

$$p_{*j} = \frac{\sum_{i=1}^{n_r} f_{ij}}{\sum_{i=1}^{n_r}\sum_{j=1}^{n_c} f_{ij}} \quad (3)$$

$$pmi_{ij} = \log\left(\frac{p_{ij}}{p_{i*}p_{*j}}\right) \quad (4)$$

$$x_{ij} = \begin{cases} pmi_{ij} & si \quad pm_{ij} > 0 \\ 0 & sinon \end{cases} \quad (5)$$

où p_{ij} est la probabilité estimée d'occurrence du mot w_i dans le contexte c_j, p_{i*} est la probabilité estimée du mot w_i et p_{*j} est la probabilité estimée du contexte c_j. Si w_i et c_j sont statistiquement indépendants, alors $p_{ij} = p_{i*} \times p_{*j}$ (par définition de l'indépendance), et ainsi pmi_{ij} est nulle (puisque $\log(1) = 0$). Le produit $p_{i*} \times p_{*j}$ est la quantité attendue si w_i se trouve dans c_j par pur hasard. D'un autre côté, s'il y a une relation sémantique intéressante entre w_i et c_j, alors d'après l'hypothèse distributionnelle, w_i et c_j co-occurrent plus souvent ensemble et donc p_{ij} est plus grand que ce qui est attendu en cas d'indépendance de w_i et c_j ; soit $p_{ij} > p_{i*} \times p_{*j}$ et pmi_{ij} est positif. Si le mot w_i n'est pas lié au contexte c_j, on doit trouver $pmi_{ij} \leq 0$. La pondération PPMI a été conçue pour donner une grande valeur à x_{ij} lorsqu'il existe une relation sémantique intéressante entre w_i et c_j ; sinon x_{ij} prend une valeur nulle, indiquant que l'occurrence de w_i dans c_j n'est pas informative. Un problème bien connu de PMI est qu'il est biaisé en faveur des événements peu fréquents. Dans le cas extrême de dépendance statistique entre w_i et c_j, on a $p_{ij} = p_{i*} = p_{*j}$ et (4) devient $-\log(p_{i*})$. PMI augmente bien lorsque la probabilité estimée du mot w_i décroit. Différents facteurs d'atténuation ont été proposés pour limiter ce problème (Pantel and Lin [2002a]).

Une autre façon de traiter les événements peu fréquents est le lissage de Laplace des probabilités estimées p_{ij}, p_{i*} et p_{*j} (Turney et al. [2003]). Une constante positive k est ajoutée aux fréquences brutes avant le calcul des probabilités : chaque f_{ij} est remplacé par $f_{ij} + k$. Plus la constante est grande, plus grand est l'effet du lissage. Le lissage de Laplace pousse les valeurs de pmi_{ij} vers 0. L'amplitude de l'effort (la

différence entre les pmi_{ij} avec et sans le lissage de Laplace) dépend des fréquences brutes f_{ij}. Si la fréquence est grande, l'effort est petit ; si la fréquence est petite, l'effort est grand. Ainsi, le lissage de Laplace réduit le biais du PMI en faveur des événements peu fréquents.

2.3.3 Lissage de la matrice

La matrice des fréquences de cooccurrence (pondérées ou non) présente deux inconvénients majeurs : elle est de grande taille et creuse. Calculer la similarité ou la distance entre paires de vecteurs est dès lors une tâche coûteuse. Diverses solutions permettent d'améliorer la complexité en temps de calcul mais aussi en espace en réduisant la dimension de l'espace de caractérisation des mots. Parmi elles :

1. ne comparer que les vecteurs qui partagent des coordonnées non nulles. La démarche consiste à calculer un index inversé des coordonnées non nulles des vecteurs de termes et à ne comparer que les vecteurs qui possèdent des composantes non nulles dont les positions coïncident. Lorsque les contextes sont des mots, des termes fréquents mais peu informatifs peuvent générer des composantes non nulles dans beaucoup de vecteurs représentatifs de mots, multipliant ainsi le nombre de vecteurs à comparer. En ne conservant que les dimensions (contexte-mot) avec un PMI au-dessus d'un certain seuil et en mettant les autres à 0, Lin [1998] a montré, dans une tâche de recherche des 200 meilleurs synonymes d'un mot, que le nombre de comparaisons nécessaires diminue fortement alors que peu de précision est perdue.

2. Deerwester et al. [1990] ont proposé une autre optimisation de la matrice terme-document, X, basée sur l'algèbre linéaire. Cette opération est la décomposition en valeurs singulières (SVD) tronquée. D'abord utilisée par Deerwester et al. dans le cadre de la similarité entre documents, Landauer and Dumais [1997] l'ont ensuite employée dans une tâche de similarité entre mots sur les QCM de synonymie du Test of English as a Foreign Language (TOEFL) et ils ont réalisé des scores proches des niveaux humains. La SVD tronquée appliquée à la similarité entre documents est appelée Latent Semantic Indexing (LSI) et elle est appelée Latent Semantic Analysis (LSA) lorsqu'elle est utilisée pour de la similarité entre mots.
   La SVD décompose la matrice pondérée X de taille $m \times n$ en produit de trois matrices $X = U\Sigma V^t$ où U, Σ et V sont de tailles $m \times m$, $m \times n$

et $n \times n$ respectivement. U et V sont deux matrices orthonormales (leurs colonnes sont orthogonales deux à deux et de longueur unitaire, $U^tU = I_m$ et $V^tV = I_n$) et Σ est une matrice dont tous les éléments sont nuls sauf sa diagonale qui est formée des valeurs singulières de X, c'est-à-dire les racines carrées des valeurs propres de la matrice X^tX (présentées dans l'ordre décroissant de leur grandeur). U est la matrice formée des vecteurs propres de XX^t et V est la matrice formée des vecteurs propres de X^tX. La SVD tronquée consiste à considérer comme approximation de X, la matrice tronquée $X_k = U_k\Sigma_kV_k^t$ où U_k, Σ_k et V_k sont des matrices de tailles respectives $m \times k$, $k \times k$ et $k \times n$. La matrice Σ_k est la matrice diagonale formée des k plus grandes valeurs singulières de X, U_k est la matrice formée des k premières colonnes de U et V_k est la matrice formée des k premières lignes de U. La matrice X_k est de rang k et c'est la matrice de rang k qui approche le mieux la matrice X, au sens où elle minimise les erreurs $\|Y - X\|_F$ sur toutes les matrices Y de rang k, avec $\|\cdot\|_F$ la norme de Frobénius (Golub and Van Loan [1996]).

Il y a différentes façons d'appréhender la SVD : recherche de sens latent, réduction de bruit, cooccurrence d'ordre élevé et réduction de disparité.

— Sens latent : Deerwester et al. [1990] et Landauer and Dumais [1997] décrivent la SVD tronquée comme une méthode d'exploration du sens latent. La SVD tronquée permet en effet de sélectionner les k principaux axes de variation de l'espace sémantique ; c'est-à-dire les k dimensions latentes les plus pertinentes. Limiter le nombre de dimensions latentes force une plus grande correspondance entre mots et contextes ; ce qui améliore la mesure de similarité.

— Réduction de bruit : Rapp [2003] décrit la SVD tronquée comme une technique de réduction de bruit. Les espaces vectoriels sémantiques présentent en effet l'inconvénient d'être bruités, du fait d'ambiguïtés telle que la polysémie, d'informations non pertinentes ou redondantes, etc. Si l'on considère que la matrice X est composée d'un mélange de signal et de bruit, avec plus de signal que de bruit, alors son approximation X_k capture essentiellement la variation de X qui est due au signal, tandis que les dimensions latentes non retenues pour la construction de X_k concernent essentiellement la variation de X qui est due au bruit.

— cooccurrence d'ordre élevé : Landauer and Dumais [1997] décrivent également la SVD comme une méthode de découverte de cooccurrence

d'ordre élevé. La cooccurrence directe (cooccurrence de premier ordre) concerne deux mots qui apparaissent dans des contextes identiques. Des cooccurrences indirectes (cooccurrence d'ordre élevé) concerne deux mots qui apparaissent dans des contextes similaires. La similarité de contextes peut être récursivement définie en termes de cooccurrence de bas niveau. Lemaire and Denhière [2006] démontrent que la SVD peut mettre en évidence des cooccurrences d'ordre élevé.

— Réduction de disparité : La matrice de cooccurrences X est généralement creuse (formée essentiellement de zéros) du fait que, même dans un grand corpus de textes, la plupart des mots sont relativement rares. La SVD tronquée permet de résoudre le problème en produisant un espace de représentation latent réduit où les mots sont représentés par des vecteurs denses. La disparité peut être vue comme un problème de manque de données : avec plus de texte, la matrice X aurait moins de zéros, et les VSM produiraient de meilleurs résultats sur la tâche choisie. Sous cette perspective, la SVD tronquée est une manière de simuler le texte manquant, en compensant le manque de données (Vozalis and Margaritis [2003]).

La SVD ne présente cependant pas que des avantages et son plus grand inconvénient est de supposer que la matrice pondérée X possède une distribution conjointe gaussienne. Minimiser la quantité $\|X_k - X\|_F$ revient à minimiser le bruit lorsque celui-ci est distribué selon une loi gaussienne, or il est connu que les fréquences de mots ne suivent pas de distribution gaussienne.

3. Depuis les travaux de Deerwester et al. [1990], beaucoup d'autres processus de lissage de matrices ont été proposés tels que la Nonnegative Matrix Factorization (NMF) (Lee and Seung [1999a]), Probabilistic Latent Semantic Indexing (PLSI) (Hofmann [1999]), Iterative Scaling (IS) (Ando [2000]), Kernel Principal Components Analysis (KPCA) (Scholkopf et al. [1997]), Latent Dirichlet Allocation (LDA) (Blei et al. [2003]), et Discrete Component Analysis (DCA) (Buntine and Jakulin [2006]).

   Ces algorithmes plus récents tendent à être plus coûteux que la SVD tronquée, mais ils sont basés sur des modèles de distributions des fréquences de mots plus réalistes que la loi normale : la loi multinomiale par exemple pour la Probabilistic Latent Semantic Indexing (PLSI) (Hofmann [1999]).

2.3.4 Comparaison des vecteurs

La manière la plus populaire de mesurer la similarité entre deux vecteurs de fréquences (brutes ou pondérées) est de calculer le cosinus de leur angle. Soient x et y deux vecteurs, avec n éléments chacun.

$$\begin{aligned} x &= (x_1, \cdots, x_n) \\ y &= (y_1, \cdots, y_n) \end{aligned}$$

Le cosinus de l'angle θ entre x et y peut être calculé comme suit :

$$\begin{aligned} \cos(x, y) &= \frac{x}{\|x\|} \cdot \frac{y}{\|y\|} \\ &= \frac{x \cdot y}{\sqrt{x \cdot x}\sqrt{y \cdot y}} \\ &= \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \end{aligned}$$

où $\cdot$ désigne le produit scalaire. Le cosinus de l'angle entre deux vecteurs est leur produit scalaire, après normalisation de leur longueur à l'unité. Notons que deux mots peuvent être synonymes tout en ayant des vecteurs représentatifs de longueurs différentes (selon qu'ils soient rares ou fréquents). Le cosinus comme mesure de similarité évite ce biais en donnant de l'importance à l'angle que forment les vecteurs entre eux et non à leurs longueurs. Le cosinus varie de -1 quand les vecteurs pointent dans des directions opposées (θ vaut 180^o) à 1 lorsqu'ils pointent dans la même direction (θ vaut 0^o). Lorsque les vecteurs sont orthogonaux (θ vaut 90^o), le cosinus est nul. Avec des fréquences brutes, qui ne peuvent avoir de valeurs négatives, le cosinus ne peut être négatif mais la pondération et le lissage de la matrice X peuvent introduire des éléments négatifs. Par exemple, la pondération PPMI ne génère pas d'éléments négatifs mais la SVD tronquée génère des éléments négatifs, même lorsque la matrice d'entrée n'en possède pas.
Une mesure de distance entre vecteurs peut facilement être convertie en une mesure de similarité par inversion (6) ou soustraction (7).

$$sim(x, y) = \frac{1}{dist(x,y)} \quad (6)$$

$$sim(x, y) = 1 - dist(x, y) \quad (7)$$

$$sim(x, y) = \exp(-dist(x, y)/\sigma) \quad (8)$$

Beaucoup de mesures de similarité ont été proposées dans la littérature (Jones and Furnas [1987] ; Lin [1998] ; Dagan et al. [1999] ; Lee and Seung [1999b] ; Weeds et al. [2004]).
En recherche d'information, il est communément admis que lorsque les vecteurs ont été correctement normalisés, les écarts de performance dus à différentes mesures de similarité sont insignifiants (Van Rijsbergen [1979]). L'usage est donc de se contenter de normaliser les vecteurs (longueur unitaire ou probabilité unitaire) avant d'utiliser n'importe quelle mesure de similarité.
Les mesures de distance les plus populaires sont les mesures géométriques comme la distance euclidienne et la distance de Manhattan ; les mesures de recouvrement comme Dice et les coefficients de Jaccard et des mesures de distance comme Hellinger, Bhattacharya, et Kullback-Leibler utilisées en théorie de l'information. Bullinaria and Levy [2007b] ont comparé ces 5 dernières mesures de distance à la mesure de similarité cosinus sur 4 tâches différentes et la mesure qui ressort comme globalement la meilleure est la mesure cosinus.
Lee and Seung [1999b] ont suggéré que les mesures qui se focalisent plus sur les caractéristiques positives des mots (coordonnées qui se recouvrent) et moins sur l'importance des caractéristiques négatives (coordonnées où un mot a une valeur non nulle et un autre mot a une valeur nulle) étaient plus performantes pour évaluer la similarité entre mots. Les mesures qui ressortent comme les meilleures dans ses expérimentations sont celles de Jaccard, Jensen-Shannon, et L1. Weeds et al. [2004] ont étudié les propriétés linguistiques et statistiques des mots similaires renvoyés par différentes mesures de similarité et ils les ont utilisées pour grouper les mesures de similarité en trois classes :

1. les mesures sensibles aux hautes fréquences (cosinus, Jensen-Shannon, α-skew, recall)
2. les mesures sensibles aux basses fréquences (precision)
3. les mesures sensibles aux fréquences similaires (Jaccard, Jaccard+MI, Lin, moyenne harmonique)

Étant donné un mot w_0, une mesure de similarité sensible aux hautes fréquences aurait tendance à donner de meilleurs scores de similarité aux mots w_i très fréquents et de moins bons scores aux mots moins fréquents. Si on utilise une mesure de similarité sensible aux basses fréquences, il y aura un biais en direction des mots à fréquence basse. Les mesures sensibles aux fréquences similaires

préféreront un mot w_i qui aura approximativement la même fréquence que w_0.
Pour finir, il n'y a pas de mesure de similarité universelle. Le choix de la mesure la plus appropriée dépend de la tâche à accomplir, de la disparité des données, de la distribution des fréquences des mots et contextes et de la méthode de lissage appliquée à la matrice.

2.3.5 Algorithmes aléatoires

Il existe des stratégies d'optimisation qui utilisent des techniques aléatoires pour approximer les mesures de similarité entre mots. Le but de ces algorithmes est d'améliorer l'efficacité calculatoire (mémoire et temps) en projetant les vecteurs à grande dimension dans des sous-espaces de dimension inférieure. La SVD tronquée agit déjà comme une projection, mais elle peut être très coûteuse et elle suppose la normalité des distributions de fréquences des mots et contextes. L'idée derrière les techniques aléatoires est que les vecteurs de grande dimension peuvent être aléatoirement projetés dans un sous-espace de dimension inférieure avec un impact relativement faible sur les scores de similarité finaux. Ravichandran et al. [1979], Gorman and Curran [2006] observent des réductions significatives dans les coûts de calcul avec en moyenne une petite erreur dans le calcul des véritables scores de similarité.
Le Random Indexing (RI) est une technique d'approximation basée sur la Sparse Distributed Memory (Kanerva [1993]). Elle calcule la similarité entre paires de vecteurs ligne d'une matrice avec une complexité $O(mnd)$, où d est une constante qui représente la taille des vecteurs index assignés à chaque colonne. La valeur de d contrôle le compromis entre précision et efficacité. Les vecteurs index sont essentiellement composés de zéros et d'un petit nombre de 1 et de -1, affectés de manière aléatoire. La mesure cosinus entre deux lignes v_1 et v_2 est approximée par le calcul du cosinus entre les deux vecteurs empreintes, empreinte(v_1) et empreinte (v_2), où empreinte(v) est le vecteur calculé en sommant les vecteurs index associés aux composantes non nulles de v. Le Random Indexing a obtenu des performances équivalentes á celles de LSA sur une tâche de sélection de synonymes de mots (Karlgren and Sahlgren [2001]).

Locality sensitive hashing (LSH) (Broder [1997]) est une autre technique qui approche la matrice de similarité avec une complexité $O(m^2 d_2)$, où d_2 est un nombre constant de projections aléatoires, qui contrôle la précision versus le compromis

d'efficacité. LSH est une classe générale de techniques pour définir des fonctions qui projettent des vecteurs (ligne ou colonne) dans des vecteurs empreintes de dimension réduite, de sorte que deux vecteurs similaires possèdent vraisemblablement des empreintes similaires. Les définitions des fonctions LSH incluent la Min-wise independent function, qui préserve la similarité Jaccard entre vecteurs (Broder [1997]), et des fonctions qui préservent la similarité cosinus entre vecteurs (Charikar [2002]). Gorman and Curran [2006] fournissent une comparaison détaillée du random indexing et de LSH sur une tâche de similarité distributionnelle. Sur le corpus BNC, LSH surpasse le Random Indexing ; cependant, sur un corpus plus large qui combine NBNC, le corpus Reuters et la majorité des English news holdings of the LDC en 2003, Random Indexing a surpassé LSH aussi bien en efficacité qu'en précision.

2.4 Notre approche pour la représentation des mots

Nous avons adopté pour nos travaux une représentation contextuelle du sens avec les modèles vectoriels sémantiques. Ce choix a été guidé par plusieurs éléments dont nos tâches d'intérêt (résumés thématiques mono et multi-documents en langue française) et les ressources réduites dont nous disposions (peu de corpus thématiques annotés en français et des ressources humaines limitées pour le travail d'annotation). Nos expérimentations ont d'abord porté sur des corpus à thème de tailles réduites puis nous les avons élargies à l'encyclopédie généraliste Wikipédia, encouragés dans cette voie par les derniers travaux qui utilisent cette ressource comme base de connaissances (Chan et al. [2013], Gabrilovich and Markovitch [2007], Hadj Taieb et al. [2013], Strube and Ponzetto [2006]). Le caractère généraliste de Wikipédia, sa taille et son évolution constante posent des difficultés classiques de mise en œuvre avec les modèles vectoriels sémantiques, comme la dimension des espaces sémantiques induits et la difficulté (voire l'impossibilité) de leur mise à jour. Pour y pallier, nous avons utilisé la technique du Random Indexing couplée avec une pondération particulière des contextes. Nous détaillons dans ce qui suit les caractéristiques de Wikipédia, la technique du Random Indexing déjà abordée plus haut et la variante pondérée que nous utilisons dans la suite de nos travaux.

2.4.1 Wikipédia comme ressource linguistique

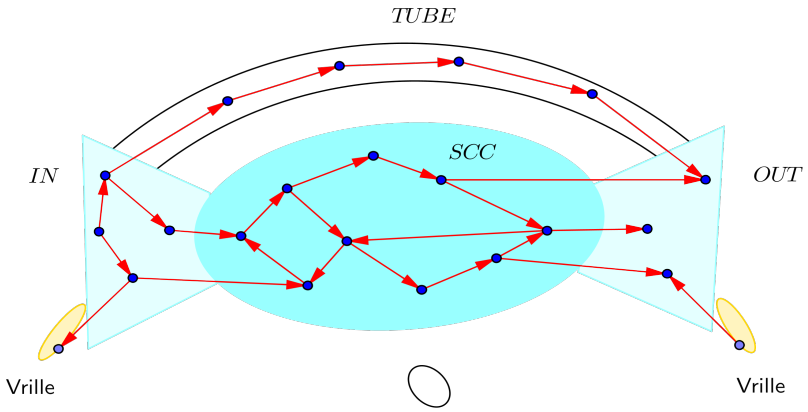


FIGURE 2.5: Structure en noeud-papillon de Wikipédia

Actuellement disponible dans 288 langues, Wikipédia est le plus grand référentiel de connaissances générales sur le Web. Les statistiques officielles de Wikipédia en date du 01/03/2015 font état de plus de 4,5 millions d'articles pour l'édition anglaise et près de 1,6 millions pour l'édition française.
L'intérêt de la communauté scientifique pour cette encyclopédie collaborative est grand et de nombreux travaux s'emploient à tirer parti de cette masse active d'informations dans des domaines divers et variés. Des études se sont plus spécifiquement attachées à décrire différentes dimensions qualitatives de Wikipédia, comme son exhaustivité, sa fiabilité, sa lisibilité et sa stabilité dans le temps. Nous renvoyons au projet Wikilit et à Mesgari et al. [2015] pour une revue détaillée de la littérature scientifique consacrée à Wikipédia depuis 2002, notamment sur des aspects comme sa structure, son contenu, ses contributeurs, ses usagers ou le corpus.

— **Structure du réseau** : Chaque article (ou concept) de Wikipédia est référencé de manière unique par une adresse URL et un article peut renvoyer à d'autres concepts via des liens hypertexte. Si l'on ne tient pas compte de la direction des liens entre articles, le graphe de Wikipédia est presque entièrement connecté : 98.5% des articles sont liés les uns aux autres. En tenant compte de la direction des liens, on retrouve la structure en nœud papillon du Web (figure 2.5) : des composantes denses fortement connectées sont liées entre elles par des liens unidirectionnels.

La zone centrale (SCC) - pour strongly connected component - est composée d'articles fortement liés entre eux : deux articles quelconques de cette zone peuvent toujours être liés par un chemin direct ou indirect. La zone (IN), de taille plus réduite, est composée d'articles qui permettent d'accéder aux articles de la zone (SCC), mais qui ne sont pas accessibles depuis cette zone. La zone (OUT), de taille équivalente à (IN) est composée à l'inverse d'articles qui sont accessibles depuis la zone (SCC), mais qui n'y renvoient pas. Les tubes sont des zones de taille plus réduites, qui relient directement les articles de la zone (IN) aux articles de la zone (OUT), sans passer par la zone (SCC). Les vrilles sont des zones atypiques qui relient des articles isolés de l'ensemble, soit à la zone (OUT), soit à la zone (IN).
Plus des 2/3 des articles de Wikipédia appartiennent au large noyau (SCC). Un signe de maturité de Wikipédia est la bonne stabilité dans le temps de ses différentes composantes ; ce qui serait actuellement le cas du Wikipédia anglais.

— **Nature sémantique des liens** : alors que dans les documents Web, un auteur peut arbitrairement lier une page à une autre, les liens dans Wikipédia indiquent une pertinence par rapport à un contexte local : un lien de la page A vers la page B indique que la page B est sémantiquement reliée au contenu, ou une partie du contenu de la page A.

— **Structure des liens** : les liens entrants dans Wikipédia ont tendance à se comporter comme les liens sortants (Jaap and Marijn [2009]) ; ce qui est consistant avec la nature sémantique des liens dans Wikipédia : si un lien de la page A vers la page B souligne une certaine pertinence de B alors il est vraisemblable que A soit également pertinent pour B.

— **Forme des articles :** La rédaction des articles de Wikipédia répond à des conventions précises. Celles-ci stipulent que le niveau de langue doit être correct, voire recherché, et exempt de faute de style ; que l'orthographe, la grammaire et la typographie doivent être irréprochables et homogènes ; que l'article doit être accessible, notamment en n'abusant pas de jargon et en renvoyant si nécessaire à des articles généralistes ou spécialisés ; que l'information doit être présentée dans une forme neutre en évitant les effets de style, les mots connotés ou chargés affectivement.

— **Domaines couverts :** Wikipédia couvre des domaines de connaissance très variés : Arts, Géographie, Histoire, Sciences, Santé, Société... Leur couverture n'est cependant pas uniforme et des catégories comme la Philosophie,

la Médecine et le Droit sont sous-représentées ; au contraire de la musique ou de la géographie qui sont sur-représentées du fait des nombreuses contributions de fans de musique ou l'insertion massive de données géographiques issues de ressources publiques. Pour exemple, en janvier 2008, la Culture et les Arts comptaient pour 30% du Wikipédia français, les biographies 15% et la Géographie 14%. La Philosophie ne représentait quant à elle que 1%. Des constatations similaires sont faites pour d'autres langues (Mesgari et al. [2015]) et l'évolution de Wikipédia dans le temps n'efface pas la sur-représentation de certains sujets et la pauvreté d'autres.

— **Fiabilité de Wikipédia :** suivant Magnus [2009], il n'est pas pertinent de parler de fiabilité globale de Wikipédia mais de fiabilité par thématique. Dans le domaine scientifique, cette encyclopédie collaborative s'avère au moins aussi précise que l'"Encyclopedia Britannica" (Giles [2005]) alors qu'elle est considérée comme peu fiable dans des domaines comme la Médecine et les Sciences humaines.

— **Évolution dans le temps :** la structure de Wikipédia et son évolution dans le temps ont été régulièrement analysés (Buriol et al. [2006], Capocci et al. [2006], Nakayama et al. [2008], Voss [2005]) et il s'avère qu'à l'instar du Web, cette encyclopédie se densifie au fil du temps aussi bien dans son contenu (nombre d'articles, longueur des articles) que dans sa structure en liens (nombre de liens entrants et sortants par article).

2.4.2 Random Indexing pondéré

Une alternative aux techniques de réduction de dimension des espaces vectoriels sémantiques repose sur l'indexation aléatoire (Random Indexing) des contextes, suivant les travaux de Pentti Kanerva sur les représentations de données éparses (Kanerva [1988], Kanerva et al. [2000]). Le Random Indexing procède en deux temps : chaque concept est représenté par un vecteur index de taille réduite puis un vecteur concept est calculé pour chaque mot par sommation des vecteurs index de tous les concepts auxquels il est associé. Les vecteurs index aléatoires sont presque orthogonaux et la technique du Random Indexing revient à projeter l'espace sémantique original dans un espace de plus petite dimension, qui préserve approximativement les distances entre points (William and Lindenstrauss [1984]). La description qui suit du Random Indexing est empruntée à Sahlgren (Sahlgren [2005a]).

On alloue un vecteur index de longueur d à chaque contexte - article Wikipédia dans notre cas. Ces vecteurs sont constitués d'un grand nombre de 0 et d'un petit nombre de 1 et de -1. À chaque composante est affectée l'une de ces valeurs avec la distribution de probabilités suivante :

$$\begin{cases} +1 & \text{avec une probabilité } s/2 \\ 0 & \text{avec une probabilité } 1-s \\ -1 & \text{avec une probabilité } s/2 \end{cases}$$

où s désigne la proportion d'éléments non nuls. Pour chaque nouveau concept, un vecteur index est produit. Dans sa version de base, le vecteur contexte d'un terme est la somme des vecteurs index de tous les contextes dans lesquels ce terme apparaît.
Ainsi, le vecteur contexte d'un terme t qui apparaît dans chacun des contextes $c_1 = [1, 0, 0, -1]$ et $c_2 = [0, 1, 0, -1]$ serait $[1, 1, 0, -2]$. Si le contexte c_1 est rencontré de nouveau, il n'y a pas création de nouveau vecteur index et la mise-à-jour du vecteur contexte de t se fait par addition du vecteur index de c_1 ; ce qui conduit au nouveau vecteur contexte de t : $[2, 1, 0, -3]$. La distance entre ces vecteurs contextes peut être évaluée au moyen de différentes mesures de distance. Sahlgren [2005b] utilise la mesure cosinus.

Une version pondérée du Random Indexing a été proposée par Gorman and Curran [2006] pour une tâche de mesure de similarité sémantique entre phrases. Le vecteur représentatif d'un mot est calculé comme la somme pondérée des vecteurs index des contextes qui lui sont associés. Les auteurs ont comparé plusieurs fonctions de pondération dans une tâche d'extraction de synonymie : fréquence du mot dans le contexte, fréquence relative, *tf-idf, tf-idf†* (version log-pondérée du tf-idf), *DICE*, etc. Ils concluent à une nette amélioration des performances de RI en présence de grands corpus de données. Pour des ensembles de données réduits, RI est suffisamment robuste et la pondération n'a, au mieux, qu'un effet mineur. Ils constatent également une grande variabilité dans l'effet des fonctions poids utilisées.
Diverses expérimentations nous ont conduits à des conclusions similaires et pour tenir compte de l'évolutivité de Wikipédia, nous avons choisi d'utiliser la fonction de pondération tf-icf (term frequency-inverse corpus frequency) qui a été introduite

par Reed et al. [2006] :

$$tf - icf_{ij} = \log(1 + f_{ij}) \times \log\left(\frac{N+1}{n_i+1}\right)$$

où f_{ij} est le nombre d'occurrences du i-ème terme dans le j-ième document, N le nombre total de documents d'un sous-corpus choisi suffisamment large et diversifié (voire le corpus Wikipédia en totalité) et n_i le nombre de documents où apparaît le terme d'indice i. Le coefficient *tf-icf* fournit une approximation du véritable *tf-idf* construit sur le corpus entier et il permet de traiter à moindre coût, des corpus dynamiques ou de très grande taille. La représentation sémantique des mots est ensuite donnée par la formule :

$$v_i = \sum_{i=1}^{n} tf - icf_{ij} \cdot c_j$$

où v_i est le vecteur représentatif du i-ème terme du vocabulaire et c_j est le vecteur index du j-ième contexte.

Nous utilisons la mesure cosinus pour évaluer la similarité entre les vecteurs représentatifs des mots.

Chapitre 3

Espace sémantique et sélection automatique des articles Wikipédia

Si Wikipédia a pour avantage la multiplicité des domaines couverts, on peut craindre en contre-partie que cette universalité puisse devenir un inconvénient lorsque l'on veut traiter d'un domaine donné. En effet, dans une encyclopédie à aussi large spectre, les termes sont utilisés dans tous leurs sens possibles et définir les termes à partir d'un corpus qui traite d'un aussi grand nombre de domaines accroît ainsi l'ambiguïté du langage. Gottron et al. [2011] avance cet argument pour expliquer les bonnes performances de la méthode ESA proposée par Gabrilovich and Markovitch [2007]. Ainsi, la première approche que nous avons testée pour construire l'espace vectoriel des termes dans un domaine donné a d'abord utilisé une étape de sélection des articles Wikipédia suivant la méthode décrite dans cette section.

3.1 Les principes

La méthode utilisée repose sur une hypothèse essentielle : dans Wikipédia, les rédacteurs qui écrivent un article dans un domaine donné proposent des liens vers les articles qui définissent les concepts de ce domaine et de ses domaines connexes. Elle est inspirée de celle utilisée par Shirakawa et al. [2009] pour la

construction d'ontologies. Au départ, le domaine concerné est défini par un ou plusieurs concepts-clefs qui correspondent à un ou plusieurs articles initiaux dans Wikipédia. Chacun de ces articles contient un certain nombre de liens vers de nouveaux concepts, qui eux-mêmes seront liés à d'autres concepts, etc. On peut donc construire un graphe de ces liens, dont les sommets sont les concepts et dont les arcs sont les liens des pages qui renvoient vers les autres pages, pondérés par la fréquence du concept dans le texte.

Certains des problèmes posés par une telle méthode sont évidents : la recherche risque de se heurter à une explosion combinatoire d'une part, et d'autre part l'intérêt des concepts ramenés par les liens ne va pas de soi. Par exemple, supposons que le domaine étudié soit le voyage d'Apollo XI sur la lune. Le concept initial est défini par l'article *Apollo XI*. Ce dernier contient 27 occurrences de *NASA* qui fait l'objet d'un seul lien, 72 occurrences de *lune* (ou *Lune*) qui fait également l'objet d'un lien ; une seule occurrence d'*océan Pacifique* qui fait l'objet d'un lien. On pressent que l'article Wikipédia traitant de l'océan Pacifique risque d'être lié à des concepts très éloignés d'Apollo XI.

Pour pallier cette double difficulté, la page pointée par un lien doit répondre à un critère de réciprocité pour appartenir à l'ensemble des concepts du domaine : les concepts d'un article ne sont retenus pour l'étape suivante que si cet article a un lien pointé vers le ou les concepts initiaux. Ainsi, l'article concernant l'océan Pacifique ne faisant pas référence à Apollo XI, il n'est pas considéré comme faisant partie des concepts du domaine.

Nous avions envisagé de restreindre à 3 la profondeur de la recherche mais, à l'expérience, le critère de réciprocité suffit à limiter l'espace des concepts d'un domaine.

3.2 Construction du Web crawler

Un concept Wikipédia est l'intitulé d'un article qui est identifié par une URL Wikipédia. Nous avons construit un Web crawler pour collectionner un sous-ensemble d'articles dans un domaine basé sur les liens et les mots-clés entrés.

La sélection des URL initiales pour le Crawler s'appuie sur le moteur de recherche de Wikipédia. Par exemple, pour un domaine définie par les mots-clefs *conquête spatiale* :

http://fr.wikipedia.org/w/index.php?search= conquête spatiale, Wikipédia va automatiquement naviguer vers l'article le plus pertinent par rapport à « conquête spatiale ». Les URL obtenues sont appelées *concepts originaux* (ou *URL originales*).

Le crawler parcourt ensuite automatiquement dans le graphe en utilisant les principes suivants :

- À partir d'une URL originale liée au domaine considéré, le crawler extrait le code-source de l'article Wikipédia ; ensuite, ce code-source est analysé pour enlever les informations inutiles (les balise HTML et XML, les URLs externes).
- Si l'article ne contient pas de liens vers les mots-clés initiaux, il n'est pas considéré.
- Si l'article contient au moins un lien vers des mot-clés initiaux, il est pris en compte :
  — ses textes et ses URLs internes sont gardés et archivés,
  — les occurrences des termes qui correspondent à ses URLs internes sont comptabilisées,
  — ses URLs internes sont utilisées pour retourner à l'étape 1.

En substance, le parcours du crawler est conforme à un algorithme de parcours en largeur du graphe. L'algorithme est le suivant :

Algorithm 1 BfsWikipedia

Input: *List-of-keyword ; URL-of-domain*
Output: *Subgraph Wikipedia*
1: **Initialization:** $queue \leftarrow URL\text{-}of\text{-}domain$
2: **While** $(queue \neq null)$
3: $link \leftarrow GetAndDrop(queue)$
4: $source\text{-}code \leftarrow GetSourceCode(link)$
5: **if** $\exists w \in List\text{-}of\text{-}keyword, w \in WordsOf(source\text{-}code)$ **then**
6: $queue \leftarrow GetLinks(source\text{-}code)$
7: $SubgraphWikipedia \leftarrow CountLinks(source\text{-}code),\ GetTexts(source\text{-}code)$
8: **End While**
9: **return** *SubgraphWikipedia*

où les procédures de calcul suivantes sont effectuées dans chaque code-source de l'article qui satisfait les conditions imposées :

- *GetAndDrop(queue)* : les URLs sont gérées par la structure *queue*, elles sont supprimées après chaque tirage
- *GetSourceCode(link)* : la procédure récupère le code-source de l'article associé à l'URL correspondante.
- *GetLinks(code-source)* : extraction des URLs internes du *code-source* pour les pousser dans la structure *queue*, chaque URL n'étant prise qu'une seule fois.
- *CountLinks(code-source)* : la procédure compte la fréquence des termes correspondant aux URL dans le *code-source*.
- *GetTexts(code-source)* : la procédure extrait le texte du *code-source*.

La figure 3.1 donne un exemple de code-source d'un article Wikipédia.

FIGURE 3.1: Le code-source d'un article Wikipédia

3.3 Calcul de la relation entre concepts Wikipédia

On note $f(c_i \rightarrow c_j)$ la fréquence du concept c_j dans l'article qui correspond au concept c_i. Le calcul sur Wikipédia au 12/12/2014 a donné les résultats suivants :

- $f(Apollo11 \rightarrow NASA) = 27$;
- $f(Apollo11 \rightarrow Lune) = 72$;
- $f(Apollo11 \rightarrow NeilArmstrong) = 35$;
- $f(NeilArmstrong \rightarrow Lune) = 30$;
- $f(NeilArmstrong \rightarrow Apollo13) = 4$;
- $f(NASA \rightarrow Lune) = 20$;
- $f(NASA \rightarrow BuzAldrin) = 1$;

Avec le sous-graphe Wikipédia obtenu, nous avons défini le poids d'un concept par rapport à un autre.

Poids des concepts : le poids $w(c_i \rightarrow c_j)$ du concept c_j dans le concept c_i est défini par la formule (3.1).

$$w(c_i \rightarrow c_j) = \frac{f(c_i \rightarrow c_j)}{\sum_{k=1}^{n} f(c_i \rightarrow c_k)}. \tag{3.1}$$

où n est le nombre de liens distincts retenus dans le concept c_i.

Par exemple, si l'on se restreint au sous-graphe de la figure 3.2 et avec les valeurs précédentes, on obtient les résultats suivants :

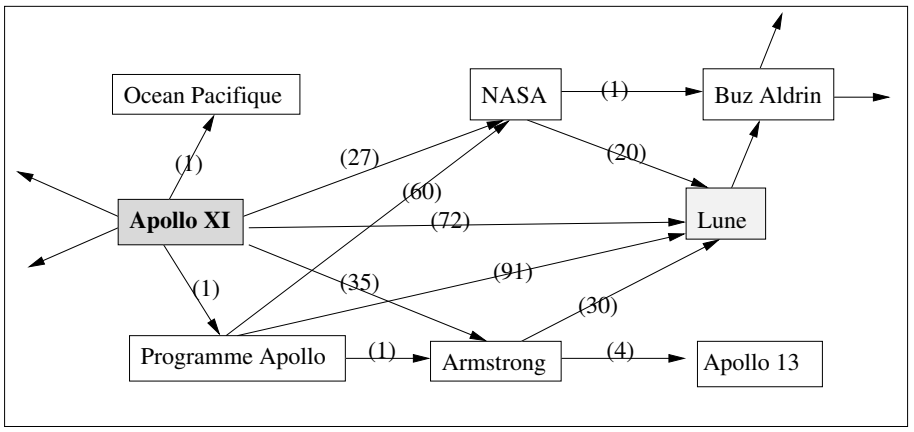
- $w(Apollo11 \rightarrow Lune) = 72/(27 + 72 + 35 + 1) = 0,53$;
- $w(Apollo11 \rightarrow NASA) = 27/(27 + 72 + 35 + 1) = 0,20$;
- $w(Apollo11 \rightarrow NeilArmtrong) = 35/(27 + 72 + 35 + 1) = 0,26$;
- $w(Apollo11 \rightarrow ProgrammeApollo) = 1/(27 + 72 + 35 + 1) = 0,01$;
- $w(NASA \rightarrow Lune) = 20/(20 + 1) = 0,95$;
- $w(NeilArmtrong \rightarrow Lune) = 30/(30 + 4) = 0,88$;
- $w(ProgrammeApollo \rightarrow Lune) = 91/(60 + 91 + 1) = 0,60$;
- $w(ProgrammeApollo \rightarrow NASA) = 60/(60 + 91 + 1) = 0,39$;
- $w(ProgrammeApollo \rightarrow NeilArmtrong) = 1/(60 + 91 + 1) = 0,01$;

Poids du chemin : $t : c_i \rightarrow c_{i+1} \rightarrow \cdots \rightarrow c_{j-1} \rightarrow c_j$ est le produit des poids de ses arcs :

$$P_t(c_i \rightarrow c_j) = \prod_{k=i}^{j-1} w(c_k \rightarrow c_{k+1}). \tag{3.2}$$

La figure 3.2 donne un exemple de sous-graphe (partiel) construit suivant les principes définis précédemment avec le calcul correspondant des chemins du concept *Apollo XI* vers le concept *Lune*.

Plus le lien entre deux concepts est long, plus le poids de ce lien est faible.



Dans cet exemple, les calculs de P_1, P_2, P_3, P_4 et P_5 correspondent à :

- $P_1(Apollo11 \rightarrow Lune) = w(Apollo11 \rightarrow NASA) * w(NASA \rightarrow Lune) = 0,20 * 0,95 = 0,19$;
- $P_2(Apollo11 \rightarrow Lune) = w(Apollo11 \rightarrow Lune) = 0,53$;
- $P_3(Apollo11 \rightarrow Lune) = w(Apollo11 \rightarrow NeilArmstrong) * w(NeilArmstrong \rightarrow Lune) = 0,26 * 0,88 = 0,23$;
- $P_4(Apollo11 \rightarrow Lune) = w(Apollo11 \rightarrow ProgrammeApollo) * w(ProgrammeApollo \rightarrow Armstrong) * w(Armstrong \rightarrow Lune)) = 0,26 * 0,01 * 0,88 \approxeq 0,00$;
- $P_5(Apollo11 \rightarrow Lune) = w(Apollo11 \rightarrow ProgrammeApollo) * w(ProgrammeApollo \rightarrow NASA) * w(NASA \rightarrow Lune) = 0,01 * 0,39 * 0,95 \approxeq 0,00$;

FIGURE 3.2: Sous-graphe partiel et exemples de calculs du poids d'un chemin.

Relation entre concepts : du fait de la densité de la structure de liaison entre concepts à l'intérieur de Wikipédia, chaque paire de concepts peut avoir un chemin direct et de nombreux chemins indirects.

La relation de c_i vers c_j, $Rel(c_i, c_j)$, est une mesure de proximité entre c_i et c_j ; elle est définie comme la somme des poids de tous les chemins de c_i vers c_j :

$$Rel(c_i, c_j) = \sum_{t}^{n} P_t(c_i \rightarrow c_j). \tag{3.3}$$

Dans l'exemple de la figure 3.2, on obtiendrait $Rel(Apollo11, Lune) = P_1 + P_2 + P_3 + P_4 + P_5 = 0,19 + 0,53 + 0,23 + 0,00 + 0,00 = 0,95$.

Avec ce mode de calcul, les paires de concepts qui n'ont pas de chemin direct obtiennent un score relativement faible. De ce fait, nous avons choisi de limiter à la valeur 3 la longueur des chemins considérés. En fait, selon nos observations, les articles d'un même domaine ont presque toujours des liens directs. Dans les deux

Épidémie	score
Pandémie	0.095292
Endémie	0.089158
Prévalence	0.082808
Épidémiologie	0.080698
Grippe aviaire	0.070050
Incidence (épidémiologie)	0.070032
Grippe	0.068509
Vaccination	0.065720
Virus de l'immunodéficience humaine	0.063178
Épizootie	0.055243
Maladie infectieuse	0.053845
Obésité	0.053440
Peste	0.052326
Organisation modiale de la santé	0.049573
Zoonose	0.042328
Grippe A (H1N1) de 2009	0.041971
Cordon sanitaire	0.040906
Charnier (tombe)	0.040837
Aéroport	0.040837
Épidémie de méningite en Afrique de l'Ouest de 2009-2010	0.040837

TABLE 3.1: Les 20 articles les plus proches du concept initial *épidémie*.

domaines que nous avons testés (cf. page 61), le nombre d'articles sélectionnés s'élève à quelques milliers.

Les tables 3.1 et 3.2 donnent les 20 articles de Wikipédia qui ont obtenu les meilleurs scores pour chacun des domaines testés, le premier défini par le concept initial *épidémie* et le second par le concept *conquête spatiale*.

Cette méthode par sélection d'articles correspond à une version spécifique de WikiRI, appelée WikiRI$_{sel}$, dont les résultats sont présentés page 64. Ils sont nettement moins bons que ceux obtenus en considérant la totalité des articles de Wikipédia, ce qui rend l'approche décevante.

Conquête spatiale	score
Mars (planète)	0.015994
États-Unis	0.014991
NASA	0.014935
Ariane (fusée)	0.012761
Vol spatial	0.012346
Union des républiques socialistes soviétiques	0.012203
Saturn (fusée)	0.012193
Module lunaire Apollo	0.011971
Navette spatiale	0.011167
Station spatiale internationale	0.010822
Skylab	0.010820
Mars 3	0.009981
Agence spatiale européenne	0.009963
Arianespace	0.009395
Liste des agences spatiales	0.009395
Programme Viking	0.009395
Atlas (fusée)	0.009395
Diamant (fusée)	0.009395
Spacelab	0.009395
Russie	0.008666

TABLE 3.2: Les 20 articles les plus proches du concept initial *conquête spatiale*.

Chapitre 4

Calculs de similarité entre phrases

4.1 Introduction

Nous avons choisi de faire reposer le calcul de la similarité entre phrases sur les vecteurs de termes. En un certain sens, ce choix restreint le champ des possibilités concernant le calcul de la similarité entre phrases. En particulier les approches compositionnelles où les termes prédicatifs sont représentés par des matrices ont été ainsi laissées de côté au profit de la robustesse et la légèreté du système. En même temps, plusieurs solutions sont possibles pour combiner les vecteurs de termes d'une phrase et pour leur adjoindre des informations syntaxiques afin d'optimiser leur utilisation. Deux approches ont été plus particulièrement étudiées.

— La première approche consiste à construire le vecteur sémantique d'une phrase à partir des vecteurs de termes qui la composent, puis à calculer la similarité entre les deux vecteurs de phrases ainsi obtenus. Dans cette approche classique, l'étude d'exemple d'associations de termes nous a conduit à proposer des pondérations qui permettent d'améliorer la méthode. Ces propositions sont décrites dans la section 4.2.

— La deuxième approche consiste à calculer la similarité entre phrases à partir des similarités de leurs termes respectifs, sans passer par la définition d'un vecteur de phrases. Par rapport à la première, cette seconde méthode permet de nombreuses expérimentations par adjonction d'informations complémentaires syntaxico-sémantiques. Nous avons mené plusieurs études qui sont décrites dans la section 4.3.

Les deux approches ont donné lieu à des évaluations en langue anglaise et en langue française dont les résultats sont présentés dans le chapitre suivant (cf. chapitre 5, page 55).

4.2 Similarité par définition d'un vecteur sémantique de phrase

Un méthode « classique » pour définir le vecteur sémantique d'une phrase par rapport aux termes qui la composent consiste à faire la somme des vecteurs de termes qui la composent (Chatterjee and Mohan [2007]), suivant la formule (4.1).

$$\vec{S} = \sum_{i=1}^{n} \overrightarrow{terme_i}. \tag{4.1}$$

Toutefois, cette mesure ne prend pas en considération le poids interne des mots dans le texte ou dans l'ensemble de textes d'où la phrase est extraite. L'hypothèse est que, si un mot est très fréquent dans les documents concernés, il convient de minimiser son importance au niveau de la phrase. Pour cela et conformément aux travaux de Neto et al. [2002, 2000], nous utilisons la pondération par le *tf-isf* (term frequency × inverse sentence frequency). Le *tf* est ici le nombre d'occurrences du terme dans la phrase et l'*isf* est calculé d'après la proportion de phrases dans l'ensemble des documents qui contiennent le terme :

$$tf\text{-}isf_{is} = tf_{is} \times \log(\frac{|S|}{SF_i}) \tag{4.2}$$

où $|S|$ est le nombre de phrases et SF_i le nombre de phrases qui contiennent le terme d'indice i. Ainsi, l'importance d'un terme qui apparaît dans un grand nombre de phrases de l'ensemble des documents s'en trouve réduite.

Par ailleurs, les vecteurs sémantiques des termes peu fréquents sont essentiellement des vecteurs creux : en d'autres termes, ils contiennent principalement des coordonnées nulles. Conformément à Higgins and Burstein [2007], les vecteurs des mots rares peuvent être enrichis en utilisant le vecteur centroïde du texte défini suivant la formule suivante.

$$\overrightarrow{centroid} = \frac{1}{n}\sum_{i=1}^{n}\overrightarrow{terme_i}, \tag{4.3}$$

où n est le nombre de termes distincts dans le texte à calculer.

Introduire dans le calcul du vecteur sémantique d'une phrase son vecteur centroïde, augmente le poids des coordonnées des vecteurs des termes rares et réduit le biais introduit par la fréquence des termes généraux. Le vecteur sémantique d'une phrase est finalement calculé avec la formule (4.4).

$$\overrightarrow{S_i} = \sum_{j=1}^{n} tf\text{-}isf_{ij} * (\overrightarrow{terme_j} - \overrightarrow{centroid}), \tag{4.4}$$

où $\overrightarrow{term_j}$ est le vecteur du terme d'indice j et n le nombre de termes distincts dans la phrase d'indice i.

4.2.1 Expérimentations concernant les groupes de deux termes et modification des pondérations

La dernière colonne de la table 4.1 contient des scores de similarité obtenus entre groupes de deux mots calculés en utilisant la méthode précédemment décrite. Les colonnes précédentes indiquent le nombre de documents Wikipédia (le *cf*) où figure chacun des mots. Dans les trois premières lignes, les deux termes associés de chaque paire sont composés d'une entité nommée désignant une personne, différente dans les deux groupes, avec un terme général, les termes généraux étant plus ou moins similaires suivant les exemples. On constate une influence qui semble a priori excessive du terme général sur le terme plus spécifique qui correspond à l'entité nommée, comme le montre particulèrement l'exemple *Rimbaud mourir/Gagarine mourir*, surtout lorsque son score est comparé à celui obtenu par *décès Rimbaud/-mort Rimbaud*.

Le premier exemple d'association entre adjectif et nom commun *petit robot/petite infection* indique une prééminence excessive de l'adjectif commun *petit* sur deux noms communs dont la similarité est très faible (0,007). Dans le dernier exemple *grande réussite/succès inattendu*, la similarité entre les noms communs est supérieure à celle des adjectifs (0,14 contre 0,09) mais les fréquences des noms

Groupes de termes	cf1a	cf1b	cf2a	cf2b	Sim
médecin Yersin/poète Rimbaud	25327	167	25845	1391	**0,124**
Rimbaud mourir/Gagarine mourir	1391	111163	481	111163	**0,926**
décès Rimbaud/mort Rimbaud	22227	1391	149554	1391	**0,348**
réussite NASA/échec NASA	8755	3528	22461	3528	**0,479**
petit robot/petite infection	166027	5930	166027	3593	**0,809**
importante réussite/grand succès	130423	8755	263987	70329	**0,447**
grande réussite/succès inattendu	263987	8755	70329	4195	**0,346**
réussite soudaine/succès inattendu	8755	4425	70329	4195	**0,172**
disparition rapide/mort soudaine	16316	32904	149554	4425	**0,217**
astronaute réussir/cosmonaute échouer	1886	41662	682	16536	**0,236**
atteindre but/réussir mission	79011	75915	41662	44349	**0,303**
astronaute piloter/médecin soigner	1886	9507	25327	8668	**0,049**

TABLE 4.1: Paires de termes : icf des termes et score de similarité WikiRI.

et des adjectifs sont très différentes : les *cf* de *grand* et d'*inattendu* sont respectivement 263 987 et 4 195 alors que ceux de *succès* et *réussite* sont 70 239 et 8 755.

Les trois exemples de paires de groupements de termes entre verbes et noms communs (dans les trois dernières lignes du tableau) donnent des résultats a priori vraisemblables.

Ces exemples semblent indiquer que lorsque l'on fait la somme d'un vecteur associé à un mot fréquent avec celui associé à un mot plus rare, le premier vecteur joue un rôle beaucoup plus important que le second. Il semble donc que, bien que l'*icf* ait considérablement réduit le poids des termes généraux, la réduction qu'il opère n'est pas suffisante : les poids des coordonnées des vecteurs termes généraux et des termes plus rares sont déséquilibrés, ceux associés aux termes fréquents étant plus importants que ceux associés aux termes moins fréquents.

L'objectif est donc de rééquilibrer le poids des termes très fréquents (mots généraux) par rapport à celui des termes plus rares, souvent spécifiques à un domaine donné, par rapport aux valeurs obtenues par le calcul classique du *tf-icf*.

4.2.1.1 Introduction du paramètre α

Pour donner toute leur importance aux termes spécifiques qui se trouvent associés à des termes généraux, on introduit un paramètre $\alpha \geq 1$, destiné à renforcer le

poids du *icf*, selon la formule (4.5).

$$tf\text{-}icf_{\alpha} = tf * icf^{\alpha}, \tag{4.5}$$

Ce paramètre diminue l'*icf* des termes tels que $icf = \log\left(\frac{N+1}{n_i+1}\right) \leq 1$, donc à très peu près ceux qui apparaissent dans plus d'un document sur dix (log désigne le logarithme en base 10) et augmente celui de ceux qui apparaissent moins souvent [1].

La figure 4.1 illustrations de l'influence du paramètre α sur les valeurs de l'*icf* d'un terme. Les différentes courbes correspondent à des valeurs de α égales à 1 (*icf* classique), 2, 3 et 4 ; chaque courbe donne la valeur de l'*icf* d'un terme en fonction du taux de documents qui le contiennent.

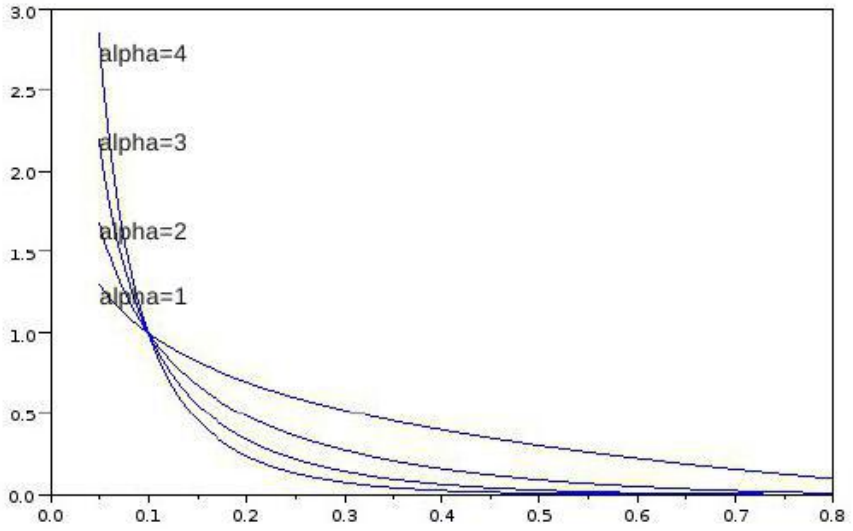


FIGURE 4.1: Valeur de $\log\left(\frac{N+1}{n_i+1}\right)^{\alpha}$ en fonction du taux de documents qui contiennent le terme pour différentes valeurs de α.

La table 4.2 donne les scores de similarités obtenus entre groupements de deux termes avec des valeurs de α égales à 1, 2, 3, 4 et 5. L'augmentation du paramètre α ramène les similarités à des valeurs intuitivement plus acceptables dans les associations entre noms communs et entités nommées : la relative rareté des entités nommées par rapport aux noms communs fait que le score de *décès Rimbaud/mort Rimbaud* augmente alors que celui de *Rimbaud mourir/Gagarine mourir* décroît. De même, le terme *petit* devient moins prépondérant dans le premier exemple d'association entre noms et adjectifs : *petit robot/petite infection*.

1. $\log\left(\frac{N+1}{n_i+1}\right) \leq 1$ ssi $\frac{N+1}{n_i+1} \leq 10$ ssi $\frac{n_i}{N} \gtrapprox \frac{1}{10}$.

Groupes de termes $\alpha =$	1	2	3	4	5
médecin Yersin/poète Rimbaud	**0,124**	0,108	0,073	0,037	0,018
Rimbaud mourir/Gagarine mourir	**0,926**	0,564	0,127	0,030	0,019
décès Rimbaud/mort Rimbaud	**0,348**	0,597	0,819	0,925	0,972
réussite NASA/échec NASA	**0,479**	0,597	0,708	0,796	0,859
petit robot/petite infection	**0,809**	0,363	0,091	0,027	0,013
importante réussite/grand succès	**0,447**	0,349	0,245	0,186	0,159
grande réussite/succès inattendu	**0,346**	0,236	0,129	0,081	0,061
réussite soudaine/succès inattendu	**0,172**	0,164	0,141	0,118	0,103
disparition rapide/mort soudaine	**0,217**	0,192	0,141	0,110	0,097
astronaute réussir/cosmonaute échouer	**0,236**	0,235	0,232	0,229	0,227
atteindre but/réussir mission	**0,303**	0,303	0,304	0,304	0,304
astronaute piloter/médecin soigner	**0,049**	0,051	0,051	0,050	0,047

TABLE 4.2: Scores de similarité WikiRI entre paires de termes associés.

Pour ce qui est des paires suivantes : *importante réussite/grand succès, etc.* l'augmentation du paramètre α fait décroître le score de similarité de manière moins facile à interpréter et il n'est pas facile d'estimer intuitivement ce que devrait être la « bonne » valeur du paramètre α.

Enfin, dans les dernières lignes du tableau, les trois exemples de paires de groupements de termes entre verbes et noms communs indiquent une grande stabilité par rapport au paramètre α et ce, malgré des valeurs très différentes de *cf* suivant les termes testés (de 1886 pour *astronaute* à 25 327 pour *médecin*).

Ces quelques exemples indiquent que le paramètre α semble avoir, dans l'ensemble, un effet bénéfique sur les scores de similarité calculés entre groupes de deux termes. D'une manière générale, son introduction permet de rétablir l'équilibre entre les mots très généraux au sens très large et les mots moins fréquents et plus spécifiques. Certains exemples montrent toutefois que le choix de la valeur optimale du paramètre n'est pas intuitivement évident. Les évaluations présentées au chapitre suivant (cf. page 55) et, en particulier celles qui concernent la langue française (cf. page 66) montrent que le choix du paramètre α est l'un des problèmes de la méthode.

4.2.1.2 Introduction de deux paramètres : α et β

Les histogrammes des figures 4.2 et 4.3 donnent respectivement le nombre de termes dans les encyclopédies Wikipédia en langue française et en langue anglaise,

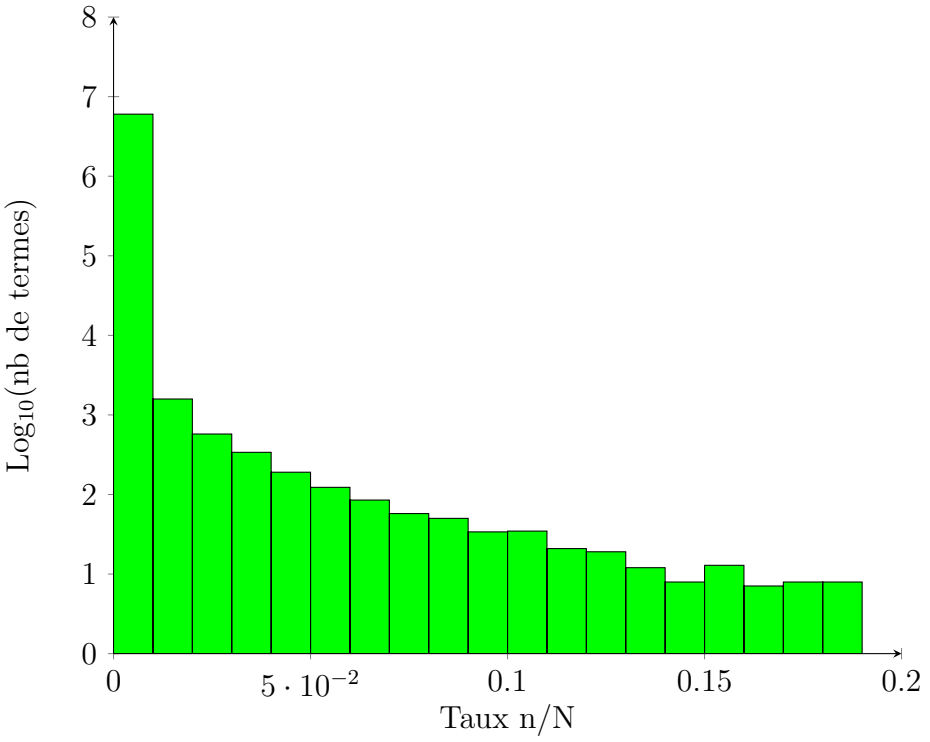


FIGURE 4.2: Logarithme décimal du nombre de termes en fonction de leur taux d'apparition dans les articles du Wikipédia français

en fonction de leur taux d'apparition dans les différents articles.

Les deux histogrammes ont des profils très similaires, même si l'on peut constater que le nombre de termes présents dans moins d'un article sur cent est proportionnellement plus important dans le Wikipédia anglais que dans le Wikipédia français. Par ailleurs, ces représentations font apparaître un phénomène auquel on pouvait s'attendre quant à la fréquence d'apparition des mots : la très grande majorité d'entre eux n'apparaît que dans un très petit nombre d'articles de l'encyclopédie.

Or, le choix du logarithme de base 10 met la barre entre les mots dont le paramètre α renforce l'*icf* et ceux pour lesquels il l'affaiblit, à environ $\frac{1}{10}$ (la valeur pour laquelle $\log\left(\frac{N+1}{ni+1}\right) = 1$). De fait, l'utilisation du paramètre α ne peut affaiblir l'*icf* que d'un très petit nombre de mots, surtout lorsque l'on considère que les mots grammaticaux sont écartés du calcul de la similarité entre phrases. Il joue malgré tout le rôle attendu dans la mesure où il renforce l'*icf* d'autant plus que cet *icf* est déjà élevé, donc que le mot est rare.

Cependant, le seuil d'un document sur dix qui sépare les « mots fréquents » des « mots rares » est un choix parfaitement arbitraire sur lequel on peut avoir envie

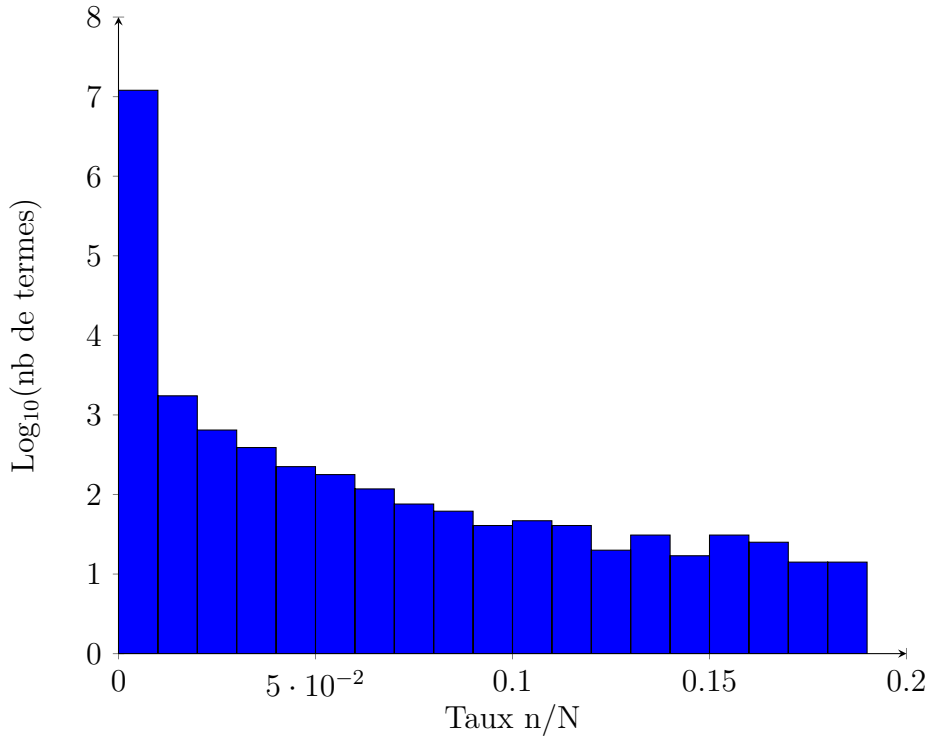


FIGURE 4.3: Logarithme décimal du nombre de termes en fonction de leur taux d'apparition dans les articles du Wikipédia anglais

d'agir. On introduit pour cela un nouveau paramètre β qui va permettre de le modifier. La définition de $icf_{\alpha\beta}$ est alors calculé suivant la formule 4.6 :

$$icf_{\alpha\beta} = \log\left(\frac{N+1}{n_i+1} - \frac{1}{\beta}\right)^{\alpha} \tag{4.6}$$

La figure 4.4 fait apparaître les seuils obtenus suivant les différentes valeurs du paramètre β. Elles correspondent aux abscisses de l'intersection des courbes avec la droite $y = 1$ (que l'on peut aisément vérifier par le calcul) : 11% pour $\beta = -1,1$, 20% pour $\beta = -0,2$, 30% pour $\beta = -0,15$ et 50% pour $\beta = -0,125$.

Les évaluations concernant l'introduction du paramètre β sont présentées page 58.

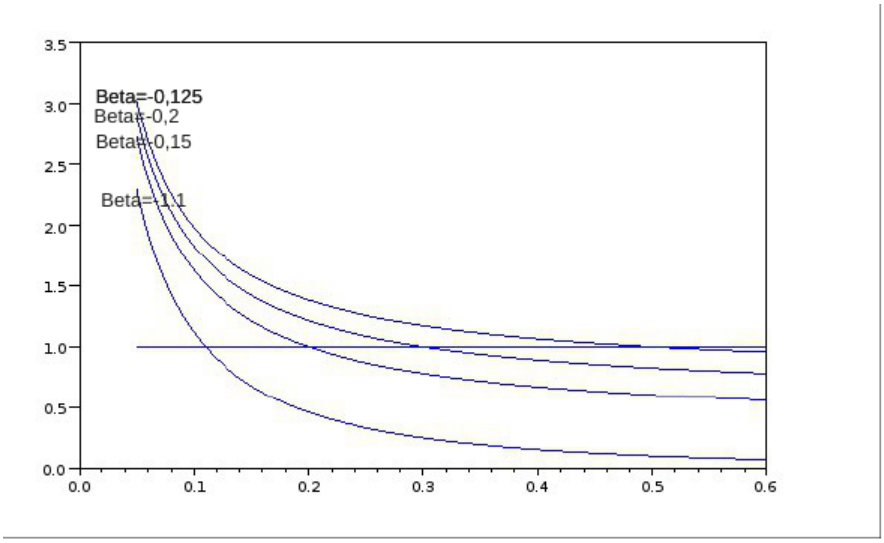


FIGURE 4.4: Valeurs de l'$icf_{\alpha,\beta}$ en fonction du taux de documents qui contiennent le terme pour différentes valeurs de β avec $\alpha = 3$.

4.3 Similarité par optimisation des similarités entre termes

Utiliser la somme des vecteurs de termes pour représenter une phrase est une méthode éprouvée qui donne des résultats acceptables : les résultats décrits dans le chapitre 5 en donneront une illustration. Cependant, il n'est pas simple d'interpréter la somme de plusieurs vecteurs sémantiques et ce caractère de « boîte noire » laisse assez peu de place aux tentatives d'amélioration.

Cette section décrit une approche qui repose sur le principe d'une similarité calculée en maximisant la somme des similarités entre les termes des deux énoncés suivant une formule proche de celle donnée par Mihalcea et al. [2006] (formule (4.7)) :

$$Sim(P_1, P_2) = \frac{1}{2}\left(\frac{\sum_{t_i \in P_1} max_{t_j \in P_2}(sim(t_i, t_j))}{n_1} + \frac{\sum_{t_j \in P_2} max_{t_i \in P_1}(sim(t_i, t_j))}{n_2}\right) \tag{4.7}$$

où n_1 est le nombre de termes de l'énoncé P_1 et n_2 est le nombre de termes de l'énoncé P_2.

Ainsi, chaque terme considéré comme sémantiquement signifiant dans chacun des énoncés est associé au terme de l'autre énoncé qui lui est sémantiquement le plus

proche au sens de la similarité entre vecteurs de termes. La formule permet d'attribuer un score compris entre 0 (similarité nulle) et 1 (similarité totale).

Un avantage de cette formule est qu'elle peut être utilisée sous diverses formes et qu'elle permet de tester l'introduction d'informations syntaxiques qui ne sont pas a priori contenues dans les vecteurs de termes. Ainsi, elle permet a priori de dépasser une simple approche « sac de mots », contrairement à ce que représente une somme de vecteurs, quelle que soit la pondération choisie.

Les différentes expérimentations que nous avons menées sont les suivantes.

1. Mihalcea *et al.* (2006) tient compte de la nature syntaxique des termes et restreint les comparaisons de similarités aux termes de même nature syntaxique : verbes, noms et adjectifs entre eux.

   Nous avons testé plusieurs types de restrictions plus ou moins contraignantes liées à la nature syntaxique des termes : de noms/noms, adjectifs/adjectifs, verbes/verbes à seulement noms propres/noms propres. Malgré ce que pouvaient laisser présager les tests de similarité entre termes en fonction de leurs natures syntaxiques, ces différents essais ont tous conduit à une détérioration très nette des résultats. On pourrait penser par exemple que restreindre une entité nommée à ne pouvoir être comparée qu'à une autre entité nommée ne peut pas détériorer les résultats mais, à l'expérience, cette discrimination induit une mauvaise similarité entre des expressions telles que « *le président japonais...* » et « *au Japon, le président...* ».

   Le système dont les résultats sont donnés dans les évaluations fonctionne donc sans restriction aucune pour ce qui est de la nature syntaxique des termes comparés.

2. Un second groupe d'expérimentations a été inspiré par les résultats négatifs de celles du premier. En effet, dans l'expression $\sum_{t_i \in P_1} max_{t_j \in P_2}(sim(t_i, t_j))$ de la formule (4.7), le même terme d'indice j peut être choisi pour plusieurs valeurs de i. Or, les expérimentations précédentes concernant les similarités entre termes ont fait apparaître que certains d'entre eux, et en particulier les verbes très courants, obtiennent des similarités relativement élevées avec un grand nombre d'autres termes, toutes catégories syntaxiques confondues. Ainsi, dans un énoncé, la présence d'un verbe général peut artificiellement augmenter son score de similarité avec les énoncés auxquels il est comparé. Nous avons donc cherché à optimiser un alignement entre les termes des deux phrases selon leur similarité qui interdise que le même terme d'un

énoncé soit utilisé plusieurs fois. Les résultats de ces expérimentations ont été, sur ce point également, nettement inférieurs à ceux obtenus avec la formule initiale. Ils ne sont pas présentés dans les évaluations du chapitre suivant.

3. Plusieurs formules de calcul de similarités entre termes ont été expérimentées. Le *cosinus* classique et le *jaccard* ont donné des résultats très similaires et nettement supérieurs aux autres formules de similarités testées. Sur l'ensemble des tests d'évaluation, ces deux mesures obtiennent des résultats qui ne diffèrent qu'à partir de la quatrième décimale : les résultats donnés dans le chapitre 5 sont donc à rapporter indifféremment à l'une ou l'autre. Sur ce sujet, il convient de remarquer que l'usage du Random Indexing tel que nous l'avons expérimenté contraint à adapter les formules de similarité entre vecteurs couramment utilisées car les positions aléatoires des coordonnées des vecteurs index égales à -1 et 1 font qu'une même coordonnée positive pour un vecteur peut être négative pour un autre.

4. Le dernier groupe des expérimentations qui ont été menées concerne la prise en compte de l'ordre des mots ou de leurs relations syntaxiques.

   Les premiers essais ont testé des adaptations de la formule (4.7) aux bigrammes des deux phrases. Les bigrammes ont été définis avec ou sans les mots grammaticaux et les similarités entre bigrammes calculées suivant diverses méthodes : moyenne des similarités optimisées terme à terme, formule (4.7) restreinte aux bigrammes, etc. Dans tous les cas, les résultats ont été inférieurs à ceux obtenus avec les unigrammes. Par ailleurs, combiner bigrammes et unigrammes n'a apporté aucune amélioration aux résultats obtenus avec les simples unigrammes.

   Une seconde série d'essais a permis de tester la prise en compte de relations syntaxiques entre les termes. Au cours des tests ainsi réalisés, le *chunking* est la seule piste qui a induit des résultats positifs. L'opération qui s'est révélée la plus efficace consiste à constituer des groupes nominaux et verbaux, qui permettent essentiellement de rattacher adjectifs et adverbes aux termes auxquels ils se rapportent. Plusieurs calculs de similarités entre groupements de termes ainsi obtenus ont été testés : par attribution d'un « vecteur de chunk » au groupement ainsi réalisé ou par combinaison directe des similarités entre les termes regroupés, la première méthode ayant finalement été retenue. Le calcul du *vecteur de chunk* dépend du type du

chunk obtenu (verbal, nominal, etc.). Par ailleurs, un élément important dans ce calcul des similarités est la prise en compte, pour ce qui concerne les groupes nominaux, de la cardinalité : singulier, pluriel ou nombre précisé par un ordinal.

Dans ce dernier cas, la similarité entre phrases repose ensuite sur l'optimisation des similarités entre les groupements. Comme dans les expérimentations concernant les termes, la prise en compte de la nature du chunk (verbal ou nominal par exemple) ou un alignement optimal des chunks en fonction de leurs similarités respectives entraînent une très nette détérioration des performances. Les résultats du chunking sont précisés dans les résultats des évaluations (cf. sections 5.1 et 5.2).

Le chapitre suivant présente les expérimentations menées pour évaluer et comparer les différentes versions du système : calcul d'un vecteur de phrases et comparaison directe entre vecteurs de termes, avec ou sans prise en compte du chunking. Bien que WikiRI ait été initialement conçu pour la langue française, la généricité de l'approche a permis de tester le système sur les corpus de tests disponibles pour la langue anglaise. Ainsi, des évaluations des différentes versions du système ont pu être effectuées dans les deux langues.

Chapitre 5

WikiRI et similarité entre phrases : évaluations

Ce chapitre est consacré aux évaluations que nous avons menées pour évaluer les différentes versions de WikiRI. Plus précisément, nous donnons les résultats des évaluations pour les versions du système suivantes :

WikiRI$_1$ désigne la version de WikiRI qui utilise la similarité *cosinus* calculée à partir des vecteurs de phrases obtenus par sommation des vecteurs des termes qui composent ces phrases.

WikiRI$_2$ désigne la version de WikiRI basée sur l'optimisation des similarités entre les termes respectifs des deux phrases testées, selon la formule (4.7).

WikiRI$_{ch}$ est la version de WikiRI qui opère un groupement des termes (*chunking*) pour former des groupes nominaux, verbaux, etc. et qui optimise la similarité entre ces groupements.

WikiRI$_{2ch}$ combine WikiRI$_2$ et WikiRI$_{ch}$ selon la proportion fixe de 80%-20%.

5.1 Évaluations du calcul de similarités entre phrases : langue anglaise

L'existence de données de tests en langue anglaise a été la raison principale pour laquelle le système WikiRI a été adapté à cette langue. En effet, depuis 2012, la

tâche STS de SemEval confronte les résultats de différents systèmes concernant la similarité entre paires de phrases, presque tous consacrés à la langue anglaise. La version 2014 de SemEval a cependant proposé une évaluation des systèmes sur des phrases en espagnol, à laquelle 9 équipes ont participé (Agirre et al. [2014]).

5.1.1 Les corpus SemEval

L'évaluation de WikiRI a été réalisée sur les données de la tâche 10 de **SemEval-2014** (Agirre et al. [2014]) qui contient 6 corpus différents à évaluer pour l'anglais.

1. **Discussion de forum** (deft-forum)

   Le corpus *deft-forum* est composé de 450 paires d'énoncés, relativement courts. Les phrases sont souvent incomplètes ou agrammaticales, avec des mots qui peuvent être mal orthographiés.

2. **Discussion de l'actualité** (deft-news)

   Le corpus *deft-news* comporte 300 paires de phrases, en général bien formées et relativement longues. Les majuscules y ont été systématiquement remplacées par des minuscules. À l'absence des majuscules près, ce corpus est particulièrement adapté à l'évaluation de WikiRI, tant pour ce qui concerne la forme des énoncés que pour le type des informations qu'ils contiennent.

3. **Titres de l'actualité** (headlines)

   Le corpus *headlines* est composé de 750 paires d'énoncés. Leur forme est celle de titres de l'actualité : de courtes phrases, souvent incomplètes. Contrairement au corpus précédent, les majuscules y ont été conservées.

4. **Descriptions d'images** (image)

   Le corpus *images* comporte 750 paires d'énoncés, généralement très courts (moins de 6 mots non grammaticaux en moyenne). Ils sont bien orthographiés et souvent réduits à un groupe nominal qui donne le sujet de l'image décrite, tels que par exemple *A jockey riding a horse.*

5. **Définitions extraites de OntoNotes et de WordNet** (OnWN)

   OnWN est un corpus composé de 750 paires d'énoncés en moyenne très courts, censés donner une définition précise et brève. Il s'agit généralement de phrases incomplètes, qui utilisent des formes très spécifiques (*the act of, the state of, etc.*).

6. **Titres et commentaires de nouvelles sur Twitter** (tweet-news)

   Le corpus *tweet-news* est composé de 750 paires de phrases, très souvent incomplètes et agrammaticales. Le corpus contient beaucoup de hashtags, de mots en majuscules ou entre guillemets, etc.

La table 5.1 présente une analyse comparative de ces 6 corpus. Les informations données sont leur nombre de mots (non grammaticaux) par phrase, leurs pourcentages d'adverbes, d'adjectifs, de noms communs, de noms propres, de verbes, ainsi que le pourcentage moyen de mots (non grammaticaux) communs entre les phrases des paires testées. Le faible pourcentage de noms propres correspond à l'absence des lettres majuscules dans *deft-news*, à leur thématique pour les corpus *images* et *OnWN*.

Par ailleurs, on peut également noter le très important pourcentage de mots qu'ont en commun les phrases des paires testées dans chacun de ces six corpus.

	Mots/Ph	ADV	ADJ	NC	NP	V	%Com./Ph
deft-news	11,8	1,9%	11,2%	33,7%	0%	14,8%	32,6%
headlines	6,3	0,7%	7,5%	25,3%	21,1%	11,6%	22,4%
images	5,8	0,4%	10,4%	30,8%	0,7%	9,5%	25,1%
OnWN	5,25	2%	6,2%	24,9%	0,2%	14,8%	25,2%
deft-forum	6,6	6%	5,6 %	16,8%	5,2%	19%	33%
tweet-news	7,4	2,2%	5,4%	18,7%	20,8%	11,1%	19%

TABLE 5.1: Analyse comparative des différents corpus de tests de SemEval.

Excepté *deft-news*, le corpus le plus intéressant pour la tâche finale pour laquelle est conçu WikiRi est *tweet-news*. La spécificité des autres corpus aurait demandé un travail spécifique que nous n'avons pas mené.

5.1.2 Étude des paramètres α et β (WikiRI$_1$)

La version du système basée sur la définition d'un vecteur sémantique de phrases, WikiRI$_1$, utilise un paramètre α dont la fonction est de renforcer l'influence des termes rares par rapport aux termes plus fréquents (cf. section 4.1). La valeur de α a été déterminée à partir des données de tests fournies pour la tâche proposée lors de la session SemEval-2012. La même valeur de $\alpha = 3$ s'est révélée être la valeur optimale pour tous les corpus de cette session et c'est donc elle qui a été retenue pour les corpus de tests de 2014.

Seuil	8%	9%	9,5%	10%	11%	20%	30%	40%	50%
β	0,4	0,667	1,9	∞	-1,1	-0,2	-0,15	-0,133	-0,125
Résultat	0,658	0,685	0,687	**0,689**	0,690	0,694	**0,695**	0,694	0,694
Nb de mots > seuil	452	390	375	349	302	98	43	24	20

TABLE 5.2: Résultats du système avec différentes valeurs du paramètre β.

5.1.2.1 Introduction du paramètre β

L'introduction d'un second paramètre β a été proposé section 4.2.1.2. L'objectif est que ce paramètre contrôle le seuil entre termes fréquents et termes plus rares. Suivant la valeur du seuil, le premier paramètre α augmente ou diminue l'*icf* d'un terme en fonction du pourcentage de documents dans lesquels ce terme apparaît.

Pour des raisons de temps de calculs, les expériences concernant le couple (α, β) ont été menées sur un seul des corpus de SemEval 2014 (*tweet-news*, cf. section 5.1) et la valeur $\alpha = 3$ a été retenue a priori. Le tableau 5.2 indique les résultats du système pour différentes valeurs du seuil et les valeurs de β correspondantes. La dernière ligne du tableau indique le nombre de mots qui dépassent le seuil correspondant dans la version anglaise de Wikipédia.

Ces résultats indiquent que les meilleurs résultats sont obtenus pour une valeur élevée du seuil, c'est-à-dire lorsqu'un très petit nombre de mots sont considérés comme « fréquents », au sens donné par le paramètre β. Par ailleurs, on constate que l'introduction du paramètre β n'a qu'une influence faible sur les résultats obtenus.

5.1.3 Résultats obtenus par les différentes versions de WikiRI sur les corpus de SemEval 2014

En 2014, 15 équipes ont participé à l'évaluation proposée par la tâche 10 de SemEval pour la langue anglaise et les résultats de 38 systèmes ont été comparés sur la base des coefficients de corrélation de Pearson avec les gold standard des corpus.

Le tableau 5.3 donne, en fonction des 6 corpus, les résultats du meilleur système participant, de la moyenne des systèmes, de la médiane, de WikiRI_1, de WikiRI_2,

Corrélations	deft-for.	deft-news	hdln	images	OnWN	tw.-news
max Sem.	0,483	0,766	0,765	0,821	0,859	0,764
moy Sem.	0,368	0,637	0,604	0,694	0,697	0,616
Médiane	0,366	0,662	0,671	0,756	0,780	0,647
WikiRI$_1$	**0,470**	0,638	**0,566**	**0,759**	0,740	0,689
WikiRI$_2$	0,430	**0,736**	0,562	0,752	**0,789**	0,720
WikiRI$_{ch}$	0,369	0,657	0,563	0,716	0,767	0,698
WikiRI$_{2ch}$	0,434	0,732	**0,567**	**0,758**	**0,788**	**0,722**

TABLE 5.3: Résultats obtenus sur les données de SemEval 2014 : corrélations obtenus par WikiRI comparées aux systèmes participants.

Rang/39	deft-for.	deft-news	hdln	images	OnWN	tw.-news	Ens.
WikiRI$_1$	15	27	30	20	23	14	23
WikiRI$_2$	**12**	**11**	30	23	20	**12**	14
WikiRI$_{ch}$	24	22	30	26	23	13	24
WikiRI$_{2ch}$	19	**11**	30	21	20	**12**	13

TABLE 5.4: Résultats obtenus sur les données de SemEval 2014 : inter-classement de WikiRI par rapport aux 38 systèmes participants.

de WikiRI$_{ch}$ et de WikiRI$_{2ch}$. Pour ce qui concerne WikiRI$_1$, les résultats sont donnés dans la cas où $\alpha = 3$, et sans introduction du paramètre β.

Le tableau 5.4 donne l'inter-classement de WikiRI dans ses différentes versions par rapport aux 38 systèmes qui ont participé à l'évaluation. La dernière colonne (*Ens.*) donne le classement de chacune des versions rapporté à l'ensemble des corpus de tests.

Les résultats de WikiRI$_1$ sont supérieurs à la médiane des systèmes participants dans 3 des 6 corpus : *deft-forum, images* et *tweet-news*. Ceux de WikiRI$_2$ sont supérieurs à la médiane des systèmes participants dans 4 des 6 corpus : *deft-forum, deft-news, OnWN* et *tweet-news*. Surtout, WikiRI$_2$ surclasse largement WikiRI$_1$ dans les deux corpus représentatifs de la tâche que sont *deft-news* et *tweet-news*. WikiRI$_{ch}$ obtient de moins bons résultats que WikiRI$_2$ mais il surclasse néanmoins WikiRI$_1$ dans ces deux corpus. Enfin WikiRI$_{2ch}$ obtient les meilleurs scores dans quatre des six corpus et il est au-dessus de la médiane des systèmes participants dans tous les corpus, excepté le corpus *headlines*. Ce dernier corpus est d'ailleurs celui où toutes les versions de WikiRI obtiennent des résultats médiocres, ce qui laisse penser que, plus encore que les cinq autres, il aurait dû faire l'objet d'un traitement ou d'un prétraitement spécifique.

Les résultats donnés dans la table 5.4 confirment les résultats précédents. Les inter-classements de WikiRI_2 et WikiRI_{2ch} parmi les 38 systèmes participants dans les deux corpus représentatifs de la tâche *deft-news* et *tweet-news* sont respectivement égaux à 11/39 et 12/39.

D'une manière générale, on peut donc dire que les résultats obtenus avec les différentes versions de WikiRI sont encourageants. Comme indiqué précédemment, les tests ont été réalisés sans utilisation de règles particulières en rapport avec les types de corpus évalués, le but n'étant pas d'optimiser les performances sur les différents corpus de SemEval, mais de valider l'approche et de pouvoir comparer les différentes versions de WikiRI. Les corpus tels que *OnWN*, *headlines* et *images* proposent des données de tests très spécifiques pour ce qui concerne la forme des énoncés. Obtenir de bons résultats dans un challenge tel que SemEval pour ce type de corpus aurait demandé la mise en œuvre de traitements particuliers. Par exemple, dans le corpus *OnWN*, la simple élimination du chunk *"the act of"* dans le calcul de similarité fait grimper le score de plusieurs points, tant la proportion d'énoncés commençant par ce groupe de mots est importante.

Par ailleurs, on peut tirer plusieurs conclusions de la comparaison entre les différentes versions de WikiRI. Si WikiRI_1 obtient le meilleur score dans le corpus *deft-forum*, il est largement distancé par WikiRI_2 dans les deux corpus les plus proches de la tâche visée que sont *deft-news* et *tweet-news*. De plus, si la version de WikiRI basée sur les similarités entre chunks n'obtient jamais les meilleurs scores, elle permet, combinée avec WikiRI_2, d'améliorer les résultats de cette version dans quatre des six corpus et elle obtient de meilleurs résultats que WikiRI_1 dans deux d'entre eux (*OnWN* et *deft-news*). Certes, une utilisation efficace d'informations syntaxiques se heurte à de nombreuses difficultés : énoncés incomplets ou agrammaticaux, mots inconnus, erreurs des parseurs, etc. En même temps, le fait de constater une amélioration des résultats lorsque l'on insère un certain nombre d'informations syntaxiques dans une approche par sacs de mots, prouve l'intérêt de la démarche. Par ailleurs, il va de soi que la méthode se doit d'être adaptée à la nature et au niveau de langue des corpus que l'on désire analyser : on ne peut pas espérer une analyse syntaxique un tant soit peu fiable d'un corpus tel que *deft-forum*.

5.2 Évaluations du calcul de similarités entre phrases : langue française

Si SemEval2014 contient des données pour l'anglais et pour l'espagnol, il n'existe pas actuellement, à notre connaissance, de corpus annoté en français pour la tâche qui nous intéresse. Les évaluations que nous avons menées ont donc été précédées de la constitution de deux corpus de taille modeste. Par ailleurs, nous n'avons pas pu comparer les résultats de WikiRI à ceux de systèmes concurrents sur ces corpus.

5.2.1 Les corpus d'évaluation

Créer un corpus de paires de phrases annotées en similarité est un travail long et difficile : tester toutes les paires d'un ensemble de n phrases devient rapidement impraticable de par la croissance quadratique du nombre de paires en fonction de n. Par ailleurs, la subjectivité de l'annotation pose également problème car elle rend la tâche d'annotation chronophage et délicate.

Nous avons extrait du Web deux corpus de textes français dans deux domaines différents définis respectivement par les mots-clefs « Épidémies » et « Conquête spatiale ». Dans chacun de ces deux corpus, nous avons sélectionné un ensemble de soixante-dix phrases, dont la longueur varie de 10 à 65 mots. Dix d'entre elles ont été choisies comme phrases de référence : elles contiennent diverses informations importantes concernant les domaines testés. Chacune de ces dix phrases a été associée à six autres phrases choisies de telle sorte que les différents niveaux de similarité entre phrases (sur une échelle de 0.0 à 4.0) soient représentés.

La table 5.5 contient les mêmes indications que celles données pour le corpus SemEval : le nombre de mots non grammaticaux par phrase, les pourcentages d'adverbes, d'adjectifs, de noms propres et de verbes ainsi que la moyenne du nombre de mots non grammaticaux communs entre les phrases des paires testées. Ces données montrent que les phrases sont notablement plus longues que celles des corpus de SemEval, exception faite de celles du corpus *deft-news*. Par ailleurs, nous avons construit nos corpus dans la perspective de l'application visée, le résumé multi-documents ; de ce fait le pourcentage de mots communs entre phrases est beaucoup plus faible que celui observé dans les différents corpus de SemEval, notre

échantillon se voulant représentatif de la tâche à laquelle devrait se confronter le système.

	Nb_Mots/Ph	ADV	ADJ	NC	**NP**	V	Com./Ph
Epid.	12,6	2,5%	10,9%	22,7%	**3,7%**	10%	9,7%
Conq. spat.	16,1	2,2%	10,7%	21,4%	**8,1%**	11,4%	6,8%

TABLE 5.5: Comparaison des corpus de tests *épidémies* et *conquête spatiale*.

La Table 5.6 donne l'une des phrases de référence du corpus *Conquête spatiale* (en gras) avec les six phrases qui lui ont été associées.

(1) ***Mars est l'astre le plus étudié du système solaire, puisque 40 missions lui ont été consacrées, qui ont confirmé la suprématie américaine - des épopées Mariner et Viking aux petits robots Spirit et Opportunity (2003 et 2004).***
(2) *Le 28 novembre 1964, la sonde Mariner 4 est lancée vers Mars, 20 jours après l'échec de Mariner 3.*
(3) *Les robots Spirit et Opportunity, lancés respectivement le 10 juin 2003 et le 8 juillet 2003 par la NASA, représentent certainement la mission la plus avancée jamais réussie sur Mars.*
(4) *Le bilan de l'exploration de Mars est d'ailleurs plutôt mitigé : deux tiers des missions ont échoué et seulement cinq des quinze tentatives d'atterrissage ont réussi (Viking 1 et 2, Mars Pathfinder et les deux MER).*
(5) *Le 6 août 2012, le rover Curiosity a atterri sur Mars avec 80 kg de matériel à son bord.*
(6) *Arrivé sur Mars en janvier 2004 comme son jumeau Spirit, et prévu comme lui pour fonctionner au moins trois mois, Opportunity (alias MER-B) roule encore et plusieurs de ses instruments répondent présents.*
(7) *Mars est mille fois plus lointaine que la Lune et son champ d'attraction plus de deux fois plus intense : la technologie n'existe pas pour envoyer un équipage vers Mars et le ramener sur Terre.*

TABLE 5.6: Les scores de similarité d'une phrase de référence avec ses six phrases associées.

Paires des phrases	(1)-(2)	(1)-(3)	(1)-(4)	(1)-(5)	(1)-(6)	(1)-(7)
Score de similarité	0,49	2,06	1,86	1,19	1,57	1,1

TABLE 5.7: Les scores de similarité de la phrase de référence de la table 5.6 avec ses six phrases associées.

Sept volontaires humains, âgés de 18 à 60 ans, ont été impliqués dans la tâche d'annotation dont trois experts et quatre candides. Ils ont évalué la similitude des

paires de phrases sur une échelle de 0,0 à 4,0 (les décimales étaient autorisées), selon les consignes indiquées dans la Table 5.8 et suivant la procédure d'annotation décrite dans Li et al. [2006].

4.0 :	*les phrases sont complètement équivalentes ;*
3.0 :	*les phrases sont globalement équivalentes, mais elles diffèrent par quelques détails ;*
2.0 :	*les phrases ne sont pas équivalentes, mais elles partagent certaines parties de l'information ;*
1.0 :	*les phrases ne sont pas équivalentes, mais elles traitent du même sujet ;*
0.0 :	*les phrases n'ont aucun lien.*

TABLE 5.8: Les instructions d'annotation pour le choix du score de similarité entre phrases

Les participants ont travaillé indépendamment et sans contrainte de temps sur une application Web [1] conçue pour leur faciliter la tâche d'annotation. Pour chaque phrase de référence choisie au hasard, ses six phrases associées ont été aléatoirement et successivement présentées à l'annotateur. Ce dernier disposait d'un historique des scores de similarité qu'il avait déjà attribués et il était libre de les modifier à tout moment. La figure 5.1 montre l'interface de l'utilitaire d'annotation dont diposaient les annotateurs.

Les données de la Table 5.7 correspondent à la moyenne des scores de similarité attribués par les sept annotateurs à chacune des six paires de phrases données dans la Table 5.6.

Pour estimer l'accord inter-annotateurs, nous avons comparé les scores de chaque annotateur à la moyenne des scores calculée sur le reste du groupe. Les coefficients de corrélation ainsi obtenus sont présentés dans la table 5.9 [2]. Compris entre 0,8 et 0,941, ils indiquent un bon accord entre les évaluateurs humains, tout au moins quant à l'échelle des similarités. Les évaluateurs sont donc largement d'accord sur les définitions utilisées dans l'échelle, même s'ils ont trouvé la tâche d'annotation particulièrement difficile.

1. http ://vuhaihieu-001-site1.smarterasp.net
2. Le choix de laisser les annotateurs utiliser des valeurs décimales ne permettait pas d'utiliser un kappa pour estimer l'accord.

FIGURE 5.1: Utilitaire d'annotation des corpus en similarité pour la langue française.

Annotateurs	1	2	3	4	5	6	7
Corrélation (c. spatiale)	0,872	0,869	0,844	**0,941**	0,886	0,815	0,855
Écart type (c. spatiale)	0,586	0,640	0,714	0,364	0,624	0,671	0,568
Corrélation (épidémies)	0,862	0,904	0,903	0,931	0,846	0,846	**0,800**
Écart-type (épidémies)	0,544	0,514	0,622	0,367	0,651	0,580	0,617

TABLE 5.9: Les coefficients de corrélation entre les scores de chaque annotateur et la moyenne des scores des six autres.

5.2.2 Résultats obtenus par les différentes versions de WikiRI sur les corpus de langue française

5.2.2.1 WikiRI sur sélection d'articles

Chacun des deux corpus français est relatif à un domaine spécifique, défini par un mot clef. Ils ont donc permis de tester la sélection d'article décrite dans la section 3.

Le tableau 5.10 donne les résultats obtenus dans chacun des deux corpus en fonction du nombre de concepts retenus.

Conquête spatiale			
nombre de concepts	1492	1857	3349
score	0,554	0,558	0,555
Épidémie			
nombre de concepts	1721	3002	5094
score	0,525	0,533	0,521

TABLE 5.10: Résultats de WikiRI avec sélection d'articles sur les corpus français (WikiRI$_{sel}$).

Ces résultats semblent faire apparaître dans les deux cas qu'il existe un nombre optimum de concepts pertinents à sélectionner et qu'au delà d'un certain seuil, utiliser des articles plus éloignés du domaine introduit du bruit dans le système.

Par ailleurs, pour chacun des deux corpus, la version WikiRI$_1$ du système a été testée avec différentes valeurs du paramètres α. Les résultats sont donnés dans la table 5.11. Alors que la valeur optimale du paramètre α reste stable entre les différents corpus en langue anglaise de SemEval, il n'en est pas de même entre les deux corpus de domaine en langue française où les valeurs optimales de α s'avèrent être très différentes : le meilleur résultat est obtenu avec $\alpha = 2,25$ pour le corpus *épidémies* et $\alpha = 4,75$ pour le corpus *conquêtes spatiales*. Enfin, l'introduction de ce paramètre s'avère très efficace voire nécessaire à l'obtention de résultats acceptables : les résultats obtenus pour $\alpha = 1$, qui correspondent à l'utilisation du *tf-icf* classique, sont largement inférieures à ceux obtenus pour les valeurs optimales : 0,648 contre 0,800 et 0,648 contre 0,849.

Ces résultats sont à comparer avec ceux de WikiRI$_{sel}$, la version de WikiRI$_1$ qui a été implémentée en limitant le nombre d'articles Wikipédia sélectionnés à ceux liés au domaine du corpus (cf. chapitre 3). Globalement, les résultats obtenus en utilisant l'ensemble du corpus Wikipédia sont nettement au-dessus de ceux obtenus en limitant l'espace des concepts : le bruit engendré par la totalité du corpus (ambiguïté accrue, etc.) semble largement compensée par les avantages d'un corpus plus important. Par ailleurs, on constate que les résultats de WikiRI$_{sel}$ sont également améliorés par l'introduction du paramètre α mais avec des valeurs optimales différentes : 1,25 pour le corpus *Epidémies* et 4 pour le corpus *Conquête spatiale*.

WikiRI$_1$ α	1	2	**2,25**	2,5	3	4,5	**4,75**	5
Epidémies	**0,648**	0,794	**0,800**	0,796	0,775	0,701	0,687	0,672
Conq. spat.	**0,648**	0,750	0,761	0,771	0,792	0,848	**0,849**	0,847
WikiRI$_{sel}$ α	1	**1,25**	1,5	2	3,75	**4**	4,25	4,5
Epidémies	**0,525**	**0,539**	0,539	0,513	0,381	0,362	0,341	0,319
Conq. Spat.	**0,556**	0,561	0,561	0,566	0,614	**0,617**	0,617	0,612

TABLE 5.11: Résultats comparés de WikiRI$_1$ et WikiRI$_{sel}$ sur les deux corpus en langue française, suivant différentes valeurs du paramètre α.

	WikiRI$_{\alpha=1}$	WikiRI$_1$	WikiRI$_2$	WikiRI$_{ch}$	WikiRI$_{2ch}$
épid.	0,648	0,800	**0,855**	0,776	0,848
conq. spatiale	0,648	0,849	0,854	0,749	**0,855**

TABLE 5.12: Résultats comparés des différentes versions de WikiRI sur les corpus en langue française.

5.2.2.2 Comparaison entre WikiRI$_1$ et WikiRI$_2$

Le tableau 5.12 permet de comparer les résultats de WikiRI$_1$ obtenus avec $\alpha = 1$, puis avec le α optimal, WikiRI$_2$, WikiRI$_{ch}$ et WikiRI$_{2ch}$. Comme pour la plupart des corpus de SemEval, les mesures de similarités faites en optimisant les similarités terme à terme obtiennent de meilleurs résultats que WikiRI$_1$, même en choisissant le meilleur α pour chacun des deux corpus.

En revanche, l'utilisation du chunking n'améliore pas les résultats de manière significative ; mais ces résultats sont déjà très élevés si on les compare à ceux obtenus avec les corpus de SemEval. Ces bons scores sont peut-être attribuables à la moyenne nettement moins élevée du nombre de mots communs entre les phrases des paires testées ; leur existence éventuelle permet en effet de détecter plus facilement la similarité entre paires d'énoncés.

5.3 Conclusion

Nous avons fait le choix d'utiliser la similarité distributionnelle pour évaluer la similarité sémantique entre phrases dans un ensemble de textes relatifs à un domaine donné. Cette similarité distributionnelle s'appuie sur Wikipédia, encyclopédie universelle, évolutive et disponible dans un grand nombre de langues. Par ailleurs,

nous avons également fait le choix de baser notre modèle vectoriel sur une définition du contexte des mots très large, puisque dans notre modèle, un contexte d'apparition d'un terme est l'article Wikipédia dans lequel ce mot apparaît.

Ces choix étant faits, deux grandes approches ont été testées.

— La première consiste à réduire le nombre d'articles Wikipédia considéré aux articles lié au domaine, en mettant en œuvre une sélection des articles à partir des liens auxquels renvoient les termes des articles.

— La seconde approche consiste à considérer l'ensemble des articles de l'encyclopédie Wikipédia, la représentation d'un terme étant alors indépendante du thème général de l'ensemble des textes étudié.

La comparaison entre les deux approches repose sur les données en langue française que nous avons construites dans deux domaines différents. Elle fait apparaître une nette supériorité de l'approche qui utlise la totalité de Wikipédia. Ce résultat infirme l'idée selon laquelle il est préférable restreindre les concepts à ceux qui sont liés au domaine pour éviter le bruit et les ambiguïtés générées par l'ensemble d'une encyclopédie aussi généraliste que Wikipédia.

Dans la deuxième approche, nous avons testé plusieurs algorithmes qui ont donné lieu à la réalisation de deux systèmes WikiRI_1 et WikiRI_2.

— Le premier système (WikiRI_1) est basé sur le principe qui consiste à représenter une phrase comme un vecteur sémantique obtenu à partir de la somme des vecteurs sémantiques de ses termes.

— Le second système (WikiRI_2) repose sur une comparaison terme à terme des vecteurs sémantiques des termes qui composent les deux phrases dont on cherche à mesurer la similarité.

Les deux systèmes ont été comparés sur des données en français et en anglais. Les résultats du système WikiRI_2 sont nettement supérieurs à ceux obtenus par WikiRI_1, ce qui semble indiquer les limites du procédé classique de la sommation des vecteurs de termes, malgré les améliorations qui nous y avons apportées.

Le chapitre suivant décrit les expérimentations menées dans une tâche de résumé, en langue anglaise et en langue française, sur la base des résultats de similarité entre phrases rendus par WikiRI.

Chapitre 6

Application de WikiRI à une tâche de résumé multi-documents

La tâche de résumé multi-documents est celle pour laquelle WikiRI a été conçu. Le système dans lequel doit s'insérer WikiRI est encore en cours de développement. Les premières expérimentations ont néanmoins été évaluées et elles font l'objet de ce chapitre.

6.1 Principes généraux

Face à la masse grandissante des documents qui sont désormais à notre disposition, le résumé automatique de textes est devenu un domaine de recherche important du Traitement Automatique des Langues ; il s'est partagé en sous-domaines spécifiques et des livres entiers lui ont été consacrés, y compris en français (Inderjeet [2001], Torres-Moreno [2011]). Notre objectif est d'implémenter un système capable d'extraire les événements les plus importants d'un ensemble de documents consacrés à un sujet donné. Il s'inscrit donc dans la perspective du résumé multi-documents, un champ particulier de la tâche du résumé automatique.

La redondance est l'un des problèmes majeurs de ce type de résumés : dès lors que l'on veut résumer plusieurs documents qui traitent du même sujet, les informations importantes ont en effet de grandes chances d'y figurer plusieurs fois. Si cette caractéristique permet précisément de repérer les éléments marquants d'un domaine donné, elle nécessite que l'approche choisie permette d'éviter

d'éventuelles répétitions. Diverses méthodes ont été expérimentées pour pallier le problème ; l'une des plus connues est celle de la pertinence marginale maximale (MMR) (Goldstein and Carbonell [1998]). Un problème connexe spécifique au résumé multi-documents est la gestion des informations contradictoires : il s'agit là d'un problème complexe que nous n'aborderons pas.

Torres-Moreno [2011] propose une taxonomie des approches utilisées pour le résumé multi-documents. L'auteur propose qu'elles soient classées en trois catégories.

— La première contient les approches basées sur la structure du document : la phrase est pondérée suivant sa position dans le texte, sa longueur, etc.
— La seconde approche consiste à utiliser une représentation vectorielle (Vector Space Model). Différentes représentations et pondérations ont été imaginées : ratio de mots-clés, utilisation du titre, similarité de la phrase et d'une requête, recouvrement des documents, etc.
— La troisième catégorie englobe les approches qui utilisent des graphes pour représenter le texte.

Pour un aperçu plus complet de cette taxonomie et des exemples de systèmes, on pourra se reporter au livre de l'auteur.

Les conférences NIST/DUC ont exploré de 2001 à 2007 les problèmes liés à la tâche du résumé multi-documents et proposé différents challenges liés à cette tâche spécifique. Par exemple, en 2006 et 2007, les résumés demandés devaient permettre de répondre à une question ou à un ensemble de questions posées sur le thème de chaque ensemble de documents à résumer (Dang [2006]). Un autre challenge proposé était celui de la capacité des mises à jour d'un résumé, grâce à un second ensemble de textes contenant de nouvelles informations.

Par ailleurs, l'évaluation des systèmes dans DUC est très élaborée ; elle combine des évaluations manuelles avec des évaluations semi-automatiques telles que Pyramid (Nenkova and Passonneau [2004]), BE (Basic Elements) (Hovy et al. [2006]) et Rouge (Lin [2004]). Ces expérimentations permettent entre autres de préciser la corrélation entre ces différentes techniques d'évaluation. Il semble que Rouge ait été beaucoup utilisé par les participants pour mettre au point leurs systèmes. Basée uniquement sur la distribution des mots entre le résumé rendu par le système et celle des résumés de référence, Rouge ne peut opérer qu'à un niveau superficiel ; de plus, des études ont permis de montrer qu'il était facile de « tromper » Rouge (Sjöbergh [2007]). Cependant, les expérimentations DUC

confirment des observations de Lin [2004] selon lesquelles cette mesure est plutôt bien corrélée avec les évaluations manuelles.

Les systèmes les mieux placés de DUC 2006 et de DUC 2007 utilisent généralement l'extraction de phrases, accompagnée de pré et post-traitements, par exemple en éliminant des portions de phrases jugées inutiles (Pingali et al. [2007], Toutanova et al. [2007]). La technique de résumé multi-documents que nous avons expérimentée repose sur cette méthode d'extraction de phrases ; l'approche consiste à construire un graphe qui représente les textes à résumer : les sommets en sont les phrases et les arcs y sont étiquetés par l'indice de similarité entre phrases calculé par WikiRI. Le score des phrases est issu du calcul de l'algorithme DivRank, une variation de PageRank a priori bien adaptée à la problématique du résumé multi-documents. Le fonctionnement de cet algorithme est précisé dans la section suivante.

6.2 Description de l'algorithme DivRank

L'algorithme PageRank fonctionne sur le principe selon lequel une page Web très visitée est une page importante. Pour appliquer cet algorithme aux résumés par extraction de phrases, une méthode classique consiste à construire un graphe dont les sommets sont les phrases et où chaque arc est pondéré par la similarité entre les deux phrases qu'il relie. PageRank permet alors de donner un poids important aux phrases qui sont « les plus visitées », en l'occurrence à celles qui possèdent le plus grand nombre de phrases voisines au sens de la similarité. Cette méthode favorise la sélection de plusieurs phrases dont les sens sont très proches ; outre le fait qu'il fait courir un risque de redondance, l'algorithme exclut celles qui sont porteuses d'informations moins fréquemment rapportées.

DivRank est un algorithme proposé par Mei et al. [2010] ; l'objectif est d'améliorer PageRank en permettant que le prestige laisse néanmoins place à la diversité ; au sens du résumé multi-documents, DivRank devrait donc permettre de choisir des phrases dont l'information peut être considérée comme importante car souvent répétée, tout en privilégiant une certaine diversité des informations.

DivRank repose sur le principe d'une marche aléatoire avec liens renforcés (*vertex reinforced random walk*) qui varie dans le temps. Le poids du sommet v à l'instant

T+1 est donné par :

$$p_{T+1}(v) = (1-\lambda)p^*(v) + \lambda \sum_{u \in V} \frac{p_0(u,v).N_T(v)}{D_T(u)} p_T(u)$$

où

- λ est un paramètre compris entre 0 et 1 ;
- $p^*(v)$ est le poids initial attribué au sommet v ;
- $D_T(u) = \sum_{v \in V} p_0(u,v).N_T(v)$ où $p_0(u,v)$ est la probabilité initiale de transition entre phrases avant renforcement ;

$$p_0(u,v) = \begin{cases} \alpha.\dfrac{w(u,v)}{deg(u)}, & \text{si } u \neq v \\ 1-\alpha, & \text{si } u = v \end{cases}$$

  où
  - $w(u,v)$ désigne la similarité entre les phrases u et v ;
  - $deg(u) = \sum_{v \in V} w(u,v)$
  - α est un paramètre choisi entre $[0,1]$.
- $p_T(u)$ est le score du sommet d'indice u à l'étape T ;
- $N_T(v)$ est le nombre de fois où le sommet v a été visité à l'étape T ; on utilise l'approximation :

$$E(N_T(v)) \sim \sum_{t=0}^{T} p_t(v).$$

Sur la base des données de la tâche 2 de DUC2004, Mei et al. [2010] rapporte pour DivRank des résultats supérieurs à ceux de PageRank, MMR (marginal maximum relevance), GH (Grasshopper), etc.

6.3 Expérimentations en langue française

Les expérimentations que nous avons menées en langue française utilisent le corpus qui a été élaboré lors du projet ANR RPM2 (de Loupy et al. [2010]). Nous utilisons l'algorithme DivRank sur le graphe des phrases dont les arcs sont pondérés avec les similarités rendues par WikiRI. Nous comparons les résultats obtenus en utilisant les deux versions principales de WikiRI : WikiRI_1 et WikiRI_2.

6.3.1 Le corpus de tests

La portion du corpus que nous avons utilisée comporte 200 documents, qui sont des articles de plusieurs journaux de la presse française, publiés entre janvier et septembre 2009. Plus précisément, elle est composée de 10 articles de presse dans chacun des 20 sujets retenus dans l'actualité du moment. Chacun de ces ensembles de documents contient en moyenne environ 5000 mots et 200 phrases. Les sujets choisis et la composition du corpus en termes de catégories grammaticales sont précisés dans de Loupy et al. [2010].

Le corpus proposé par RPM2 contient également 10 autres articles par sujet, qui contiennent de nouvelles informations par rapport aux 10 articles initiaux. Cette partie du corpus a été élaborée pour pouvoir juger de la capacité d'un système à actualiser les informations. Nous ne l'avons pas utilisée. Pour la partie qui nous intéresse, le corpus RPM2 propose également quatre résumés d'une centaine de mots pour chacun des 20 thèmes retenus, élaborés par quatre annotateurs.

Nous avons évalué les résumés obtenus par notre système en utilisant pour le graphe initial les similarités calculées à partir de WikiRI_1 et WikiRI_2. Très classiquement, nous avons utilisé une version de ROUGE (ROUGE-SU2) pour évaluer la qualité des résumés obtenus. ROUGE-SU2 est basée sur la distribution des mots entre le résumé rendu par le système et celle des résumé de référence, avec possibilité de trous d'au plus deux mots dans les bigrammes.

Nous avons utilisé cette même mesure de ROUGE pour mesurer l'accord entre le résumé de chacun des quatre annotateurs avec les résumés des trois autres. Ces calculs nous ont donné les résultats indiqués dans la table 6.1 pour chacun des vingt thèmes. La moyenne des accords par thème est indiquée dans la dernière colonne, celle par annotateur dans la dernière ligne.

Comme c'est souvent le cas pour ce type de tâches, on constate que les accords inter-annotateurs donnés par ROUGE-SU2 sont faibles : la moyenne générale des accords est de 0,2378. Ils sont en même temps relativement homogènes : l'écart-type est de 0,040 et les 80 accords ainsi calculés sont compris entre 0,145 et 0,326. Lorsque l'on regarde la moyenne des accords par sujet, les résumés les moins consensuels concernent la présidence d'Obama alors que le meilleur accord est obtenu sur l'affaire du petit Mohamed (enfant perdu). Concernant ces chiffres, il convient de remarquer que les résumés produits par les annotateurs ne sont pas

	Sujet	Annot1	Annot2	Annot3	Annot4	Moy
01	Ingrid Bétancourt	0,277	0,279	0,229	0,200	0,246
02	Caisse d'Epargne	0,284	0,261	0,264	0,259	0,267
03	Crise bancaire	0,222	0,219	0,216	0,190	0,212
04	Dalaï Lama	0,235	0,215	0,213	0,181	0,211
05	Fichier Edvige	0,258	0,296	0,251	0,303	0,277
06	JO de Pékin	0,196	0,179	0,229	0,162	0,192
07	Jérôme Kerviel	0,311	0,262	0,252	0,200	0,256
08	Lance Armstrong	0,296	0,324	0,295	0,278	0,298
09	La loi Leonetti	0,214	0,225	0,211	0,186	0,209
10	Le petit Mohamed	0,259	0,319	**0,326**	0,295	**0,300**
11	Obama président	0,207	0,182	0,180	**0,145**	**0,179**
12	Licenciement de PPDA	0,284	0,312	0,293	0,292	0,295
13	Le temple de Preah Vihear	0,236	0,256	0,248	0,253	0,248
14	Election au PS	0,229	0,243	0,228	0,204	0,226
15	Grossesse Rachida Dati	0,234	0,244	0,220	0,252	0,238
16	Rachida Dati et les magistrats	0,222	0,237	0,244	0,206	0,227
17	Réforme du lycée	0,218	0,206	0,197	0,192	0,203
18	Réforme de l'audiovisuel public	0,208	0,235	0,245	0,228	0,229
19	Relance de l'économie	0,225	0,259	0,236	0,218	0,235
20	Crise au Tibet	0,177	0,215	0,210	0,230	0,208
	Moyenne	0,240	0,248	0,239	0,224	

TABLE 6.1: Évaluation ROUGE-SU2 du résumé de chaque annotateur en fonction des résumés des trois autres.

de simples extractions du corpus mais des reformulations ; ce qui peut peut-être expliquer les faibles scores rendus par ROUGE qui calcule un score de similarité en comptabilisant les mots et bigrammes communs.

6.3.2 Les résultats

Nous avons fait tourner l'algorithme du DivRank avec les deux paramètres α et λ empiriquement fixés respectivement à 0,5 et 0,7. Les arcs du graphe ont été pondérés avec les résultats rendus par WikiRI_1 puis avec ceux rendus par WikiRI_2.

La table 6.2 donne les scores des résumés ainsi obtenus par rapport aux résumés de référence des annotateurs, calculés en utilisant ROUGE-SU2.

Les résultats obtenus en utilisant les similarités rendues par WikiRI_2 sont supérieurs à ceux obtenus en utilisant celles données par $\text{WikiRI}_{1\alpha}$ dans 14 des 20 thèmes du corpus. Par ailleurs, leur moyenne, calculée sur l'ensemble des 20 thèmes, montre également une nette supériorité de WikiRI_2 sur WikiRI_1. De fait,

	Sujet	$WikiRI_1$	$WikiRI_2$	Moy_annot
01	Ingrid Bétancourt	0,165	0,132	0,246
02	Caisse d'Epargne	0,190	0,167	0,267
03	Crise bancaire	0,127	0,122	0,212
04	Dalaï Lama	0,136	0,188	0,211
05	Fichier Edvige	0,230	0,382	0,277
06	JO de Pékin	0,102	0,156	0,192
07	Jérôme Kerviel	0,164	0,251	0,256
08	Lance Armstrong	0,209	0,180	0,298
09	La loi Leonetti	0,114	0,202	0,209
10	Le petit Mohamed	0,161	0,218	0,300
11	Obama président	0,136	0,153	0,179
12	Licenciement de PPDA	0,262	0,201	0,295
13	Le temple de Preah Vihear	0,172	0,254	0,248
14	Election au PS	0,151	0,191	0,226
15	Grossesse Rachida Dati	0,242	0,239	0,238
16	Rachida Dati et les magistrats	0,149	0,162	0,227
17	Réforme du lycée	0,126	0,211	0,203
18	Réforme de l'audiovisuel public	0,114	0,162	0,229
19	Relance de l'économie	0,111	0,193	0,235
20	Crise au Tibet	0,131	0,172	0,208
	Moyenne	0,1596	0,1968	0,2378

TABLE 6.2: Scores rendus par ROUGE-SU2 pour les résumés du corpus RPM2 à partir des similarités rendues par $WikiRI_1$ et $WikiRI_2$ et en utilisant DivRank.

$WikiRI_2$ permet d'obtenir un score qui se situe à distance quasi égale de celui obtenu avec $WikiRI_1$ et de la moyenne de ceux des annotateurs.

Ces constations confirme la supériorité de $WikiRI_2$ sur $WikiRI_1$ lors des évaluations directes de ces deux systèmes. Elles démontrent également l'importance que revêt la qualité des résultats de similarité pour mettre en œuvre une approche de résumé par extraction de phrases telle que celle que nous avons choisie.

6.4 Expérimentations en langue anglaise

Nous avons expérimenté l'algorithme DivRank avec les similarités rendues par $WikiRI_1$ sur les données de DUC 2007.

6.4.1 Les données de test

Les documents à résumer dans DUC 2007 sont des articles de journaux extraits des *Associated Press*, du *New York Times* (1998-2000) et de la *Xinhua News Agency* (1996-2000). Il y a 25 documents pour chacun des 45 thèmes retenus. La tâche consiste à faire un résumé d'au plus 250 mots par thème. Au-delà, le résumé proposé est automatiquement tronqué. La table 6.3 indique, pour chaque ensemble de documents, son thème, son nombre de phrases et leur longueur moyenne.

Comme dans le cas des données de RPM2, DUC 2007 propose des documents pour évaluer la tâche de mise à jour qui concerne un sous-ensemble des 10 thèmes de la tâche de résumé. La mise à jour se fait sur la base de 25 documents supplémentaires par sujet traité et doit comporter au plus 100 mots. Une fois de plus, nous n'avons pas abordé cette tâche de mise à jour.

Par ailleurs, quatre résumés de référence ont été élaborés pour chacun des 45 thèmes mais, contrairement aux données du corpus RPM2, ces résumés ne sont pas disponibles. En revanche, DUC met à disposition un ROUGE-*BE package* pour une évaluation automatique de résumés produits.

6.4.2 Les résultats de WikiRI$_1$

Les expérimentations que nous avons effectuées ont utilisé WikiRI$_1$ avec les deux paramètres α et λ de l'algorithme du DivRank empiriquement fixés respectivement à 0,5 et 0,7.

Les résultats donnés par le paquet ROUGE fourni par DUC figurent dans la table 6.4 ; ils n'ont qu'un caractère indicatif car, comme expliqué auparavant, les résultats officiels du challenge DUC sont calculés par diverses méthodes qui associent des évaluations manuelles à des évaluations semi-automatiques. La table donne les scores respectifs du meilleur et du moins bon système, ainsi que le score médian. Le rang correspond à l'interclassement de notre système parmi les 32 systèmes participants.

Comme dans le challenge SemEval, le score du système est au-dessus de la médiane des systèmes participants. Aucun pré ou post-traitement n'a été implémenté malgré l'importance qu'ils peuvent présenter pour améliorer les résultats. On peut

No	Thème	Nb. de ph.	Lg. moy.
1	Southern Poverty Law Center	1031	25,25
2	art and music in public schools	767	25,72
3	steps toward introduction of the Euro	155	33,90
4	Amnesty International	247	31,65
5	Basque separatism	314	28,85
6	Burma government change 1988	441	26,98
7	Turkey and the European Union	210	29,12
8	world-wide chronic potable water shortages	175	28,62
9	Angelina Jolie	821	24,12
10	Israel / Mossad "The Cyprus Affair"	324	27,07
11	Microsoft's antitrust problems	479	29,15
12	Salman Rushdie	235	29,37
13	Pakistan and the Nuclear Non-Proliferation Treaty	274	30,02
14	Napster	804	26,28
15	International Land Mine Ban Treaty	301	28,34
16	Jabiluka Uranium Mine	280	29,43
17	fen-phen lawsuits	519	25,45
18	Starbucks Coffee	604	24,85
19	unemployment in France in the 1990s	222	31,45
20	Oslo Accords	325	31,27
21	Matthew Shepard's death	621	27,10
22	US missile defense system	270	29,18
23	Senator Dianne Feinstein	795	25,44
24	obesity in the United States	740	26,22
25	Iran's nuclear capability	351	27,96
26	Al Gore's 2000 Presidential campaign	454	27,46
27	Newt Gingrich's divorce	550	24,58
28	Interferon	739	24,36
29	Eric Rudolph	560	26,01
30	line item veto	582	26,35
31	Linda Tripp	523	23,62
32	Kenya education developments	152	31,91
33	public programs at Library of Congress	770	26,90
34	acupuncture treatment in U.S.	738	24,12
35	reintroduction program for wolves in U.S.	918	26,56
36	Oprah Winfrey TV show	638	24,19
37	deep water exploration	427	28,29
38	mining in South America	233	27,50
39	after "Seinfeld"	831	23,20
40	round-the-world balloon flight	390	28,60
41	day trader killing spree	1384	22,03
42	John F. Kennedy, Jr., dies in plane crash	701	27,91
43	earthquakes in Western Turkey in August 1999	525	26,35
44	organic food	665	25,26
45	OJ Simpson developments	447	24,70

TABLE 6.3: Données concernant le corpus DUC 2007.

version ROUGE	**-1**	**-2**	**-3**	**-4**	**-L**	**-W-1-2**	**SU4**
Max	0,427	0,113	0,038	0,023	0,393	0,149	0,166
Min	0,279	0,037	0,010	0,001	0,235	0,086	0,083
Mediane	0,397	0,089	0,027	0,012	0,365	0,138	0,146
WikiRI_1 p=0,85	0,399	0,092	0,030	0,016	0,366	0,139	0,146
rang (/33)	16	11	11	6	17	15	16

TABLE 6.4: Résultats du système sur les données DUC 2007.

donc considérer que ces premiers résultats encouragent à poursuivre le travail commencé.

Par ailleurs, nous n'avons pas pu, faute de temps, obtenir les résultats de WikiRI_2. L'algorithme mis en œuvre dans WikiRI_2 est en effet plus coûteux en temps que celui mis en œuvre dans WikiRI_1. Ce défaut s'était révélé peu sensible dans les précédents tests de similarité ou de résumé. Mais, comme on peut le voir dans la table 6.3, la taille des documents de tests proposées dans DUC 2007 est notablement plus importante que celle des données de tests du corpus RPM2 et cette différence suffit à rendre WikiRI_2 peu opérationnel avec les moyens de calcul dont nous disposions.

6.5 Conclusion

Les travaux entrepris sur le résumé multi-documents ne sont pas entièrement finalisés, faute de temps. Cependant, ils offrent des perspectives intéressantes pour des travaux de recherche ultérieurs. Par ailleurs, les premiers résultats obtenus sont instructifs et positifs. La comparaison des résultats de WikiRI_1 et de WikiRI_2 sur le corpus RPM2 montrent l'importance de la tâche de similarité en sous-tâche de celle de résumé automatique lorsque l'on veut s'appuyer sur un algorithme de type PageRank. En outre, le niveau de résultat obtenu à partir de WikiRI_1 sur les données de DUC 2007 est tout à fait satisfaisant et, compte tenu de l'absence d'optimisation, ils prouvent la validité de l'approche pour réaliser un système de résumé multi-documents léger et robuste.

Chapitre 7

Bilan et perspectives

Les travaux présentés dans cette thèse n'ont pas été effectués selon un cheminement linéaire mais sont eux-mêmes le résultat de différentes expériences. Le parcours qui a été suivi mérite de figurer dans le bilan final que l'on peut tirer de l'ensemble des résultats précédemment décrits.

7.1 Objectifs initiaux et déroulement des travaux

L'objectif initial était le résumé multi-document lié à un sujet donné. Plus précisément, l'ambition du projet était de dégager la chronologie des faits marquants du domaine concerné à partir d'un choix de documents extraits du Web. Le système devait pouvoir utiliser des textes de la langue française en priorité.

Le travail s'est donc orienté vers une détermination des concepts du domaine telle qu'elle a été décrite dans le chapitre 3 et vers un système permettant de dégager les phrases les plus importantes d'un document puis d'un ensemble de documents.

L'approche choisie pour attribuer un score aux phrases était d'utiliser un algorithme du style PageRank. Très rapidement, la similarité entre phrases est apparue comme une étape préliminaire incontournable. En effet, les premiers résultats s'étant avérés très décevants, leur analyse a rapidement montré que l'algorithme s'appuyait sur un graphe où les arcs entre phrases étaient pondérés par des valeurs de similarité qui n'étaient pas cohérentes aux yeux d'un expert humain.

Les travaux se sont donc réorientés vers l'étude de la similarité entre mots, puis entre phrases. Si chacun de ces objectifs constitue déjà en soi un sujet de recherche à part entière, nous étions convenus que le système devait rester conçu en fonction de l'objectif initial du résumé multi-document ; par ailleurs, il devait pouvoir être appliqué à la langue française ou à d'autres langues moins bien dotées en ressources que ne l'est la langue anglaise. Le choix de Wikipédia comme corpus de base pour construire la représentation des vecteurs a été guidé par ces principes.

Par la suite, la construction de la représentation vectorielle des termes de la langue anglaise a été avant tout guidé par l'existence de données de tests pour l'anglais : SemEval pour tester ce qui était devenu notre objectif principal, à savoir mesurer la similarité entre phrases et DUC, pour pouvoir malgré tout tester le résumé automatique multi-documents et justifier nos choix initiaux.

7.2 Bilan

La première conclusion que l'on peut tirer des résultats concernant le résumé automatique est qu'ils justifient le choix qui a été fait de travailler en priorité sur les mesures de similarité. D'autres approches sont évidemment possibles mais, avec le choix qui avait été fait d'utiliser un algorithme tel que PageRank ou DivRank, la qualité des résultats obtenus dépend directement de ceux qui mesurent la similarité entre phrases. Les résultats comparés des deux versions décrites du système, WikiRI_1 et WikiRI_2 sur les données du corpus RPM2, illustrent bien cette remarque qui sonne comme une évidence : la supériorité de WikiRI_2 pour mesurer la similarité entre phrases sur le corpus SemEval se retrouve dans les résultats obtenus sur les résumés proposés dans RPM2.

Un autre point positif qui mérite d'être souligné est que WikiRI_1 obtient des résultats au niveau de l'état de l'art, aussi bien dans la tâche de mesure de similarité (SemEval) que dans celle de résumé automatique multi-documents (DUC). En ce sens, l'objectif est atteint de réaliser un système simple, robuste et qui ne nécessite comme ressources lexicales qu'un étiqueteur morpho-syntaxique et une version suffisamment développée de Wikipédia. Par ailleurs, l'utilisation du Random Indexing a permis d'implémenter des traitements efficaces, sans perte manifeste d'informations.

Les deux versions de WikiRI s'appuient sur la même vectorisation où un terme est représenté par un vecteur qui comptabilise ses occurrences dans chacun des articles de Wikipédia. La vectorisation que nous avons réalisée en sélectionnant les articles liés au domaine a donné des résultats nettement inférieurs. Dans les expérimentations que nous avons menées, l'avantage que représente l'utilisation d'un gros corpus l'a emporté sur le problème de l'ambiguïté introduite par son universalité.

La comparaison entre WikiRI_1 et WikiRI_2 appelle quant à elle plusieurs remarques.

— La représentation d'une phrase par un vecteur obtenu en sommant les vecteurs de ses termes est sujet à polémique. Il est vrai que cette opération est difficile à interpréter et à contrôler, comme le prouvent les problèmes rencontrés autour de la définition du paramètre α pour la langue française. En même temps, on est bien obligé de reconnaître qu'elle donne des résultats, sinon remarquables, du moins acceptables. Sur ce point-là cependant, la supériorité de WikiRI_2 (où cette sommation n'est pas mise en œuvre) sur WikiRI_1 prouve que la méthode n'est pas optimale.

— L'introduction du vecteur centroïde dans la représentation par sommation est un autre point qu'il est difficile de justifier théoriquement ; pourtant, il est indéniable que son introduction fait gagner plusieurs points de score aux résultats produits par WikiRI_1.

— La comparaison des phrases terme à terme introduite dans WikiRI_2 obtient incontestablement de bons résultats. Il est cependant troublant que tous les efforts faits pour améliorer l'approche « sac de mots » par association de termes n'ait abouti qu'à des résultats décevants. Dans cette représentation vectorielle des termes en fonction de leur contexte qui constitue le fondement de l'approche distributionnelle, le problème majeur semble bien être celui du passage du mot au groupe de mots ; comment combiner les vecteurs de termes pour obtenir une représentation vectorielle des groupes de mots que ces termes constituent ?

7.3 Pistes d'amélioration et perspectives

Une piste d'amélioration, radicale, serait de modifier la représentation des termes. Ainsi, pour mieux évaluer la similarité entre groupes de mots ou phrases, les

modèles enrichis essaient d'hybrider des modèles que Grefenstette and Sadrzadeh [2011] qualifie d'orthogonaux, à savoir les modèles distributionnels avec ceux de la sémantique formelle : grammaires catégorielles, grammaires de prégroupes de Lambek, etc.

Ces approches permettent ainsi de prendre en compte l'ordre des mots et/ou les relations syntaxiques qui les réunissent en s'appuyant sur le caractère compositionnel de la langue : représentation des adjectifs par des matrices qui agissent sur le vecteur qui représente le nom qu'ils qualifient (Baroni and Zamparelli [2010]), matrice de verbes agissant sur les sujets et compléments d'objets (Coecke et al. [2010], Grefenstette and Sadrzadeh [2011]), représentation des phrases comme un arbre de vecteurs (Socher et al. [2011]), etc. Une autre possibilité consisterait à utiliser les modèles de langage en réseaux de neurones (NNLM) ; l'un des objectifs des propositions de Mikolov et al. [2013a] est de réduire les coûts d'apprentissage de ces modèles, par ailleurs très efficaces.

Mis à part le problème fondamental de la représentation d'un groupe de mots à partir de celles des mots qui le constituent, plusieurs pistes d'amélioration, plus modestes, sont possibles.

— Une première amélioration possible concerne la vectorisation des termes. Il faudrait entre autres améliorer le prétraitement du corpus (gestion des majuscules par exemple) et corriger certains bugs introduits par Treetagger. Un autre progrès important serait celui de la gestion des entités nommées et de leur identification sous plusieurs formes ; par exemple actuellement « Henri Matisse » est comptabilisé dans la phrase comme la juxtaposition de deux termes indépendants : « Henri » et « Matisse ».
Une autre piste serait d'utiliser un modèle tel que GloVe (cf. page 9).

— L'introduction du chunking dans WikiRI$_2$ a produit une très légère amélioration des résultats mais ses effets sont quand même décevants par rapport à l'alourdissement des traitements qu'il impose. Il faudrait probablement effectuer un étiquetage syntaxico-sémantique du corpus, du type de celui qui est fait dans le système RITEL (Rosset et al. [2008]) et travailler sur le choix de la représentation du chunk en fonction de son étiquetage.

— Une autre piste serait d'explorer d'autres mesures de similarités inter-phrases telles que celle basée sur un « produit scalaire élastique » qui permet de prendre en compte le contexte d'occurrence des mots en considérant leur ordre.

Les pistes sont nombreuses et le problème se pose de préserver à la fois la robustesse et la généricité du système.

Annexe A

Liste des publications

1) Hai-Hieu Vu, J. Villaneau, F. Saïd and P-F. Marteau. "Mesurer la similarité entre phrases grâce à Wikipédia en utilisant une indexation aléatoire". TALN2015. ATALA, 2015.

2) Hai-Hieu Vu, J. Villaneau, F. Saïd and P-F. Marteau. "Sentence Similarity by Combining Explicit Semantic Analysis and Overlapping N-Grams". TSD. Springer International Publishing, 2014.

3) Hai-Hieu Vu, J. Villaneau, F. Saïd. "Utilisation des liens wikipedia pour la détection automatique des concepts d'un domaine". CLIF, 2013.

Bibliographie

Acar, E. and Yener, B. (2009). Unsupervised multiway data analysis : A literature survey. *IEEE Transactions on Knowledge and Data Engineering*, 21 :6–20.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., and Wiebe, J. (2014). Semeval-2014 task 10 : Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Ando, R. K. (2000). Latent semantic space : Iterative scaling improves precision of interdocument similarity measurement. In *Proceedings of the 23rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2000)*, pages 216–223.

Balasubramanian, N., Allan, J., and Croft, W. B. (2007). A comparison of sentence retrieval techniques. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 813–814. ACM.

Baroni, M. and Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices : Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1183–1193, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bengio, R., Ducharme, P., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3 :1137–1155.

Benzécri, J.-P. (1980). L'analyse des données : l'analyse des correspondances. Bordas, Paris.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3 :993–1022.

Broder, A. (1997). On the resemblance and containment of documents. In *In Compression and Complexity of Sequences (SEQUENCES'97)*, pages 21–298.

Budanitsky, A. and Hirst, G. (2001). Semantic distance in wordnet : An experimental, application-oriented evaluation of five measures. In *Proceedings of the Workshop on WordNet and Other Lexical Resources. Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001).*

Bullinaria, J. and Levy, J. (2007a). Dirt-discovery of inference rules from text. *Extracting semantic representations from word cooccurrence statistics : A computational study*, 39 :510–526.

Bullinaria, J. and Levy, J. (2007b). Extracting semantic representations from word cooccurrence statistics : A computational study. In *Behavior Research Methods*, volume 39, pages 510–526.

Buntine, W. and Jakulin, A. (2006). Discrete component analysis. In *Subspace, Latent Structure and Feature Selection : Statistical and Optimization Perspectives Workshop at SLSFS 2005*, pages 1–33.

Burgess, C. and Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. In *Language and Cognitive Processes*, volume 12, pages 177–210.

Buriol, L. S., Castillo, C., D., D., S., L., and S., M. (2006). Temporal analysis of the wikigraph. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference of Web Intelligence.*, pages 45–51.

Capocci, A., Servedio, V., Colaiori, F., and Buriol, L. (2006). Preferential attachment in the growth of social networks : the case of wikipedia. *Arxiv preprint physics.*

Chan, P., Hijikata, Y., and Nishida, S. (2013). Computing semantic relatedness using word frequency and layout information of wikipedia. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 282–287. ACM.

Charikar, M. S. (2002). Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing (STOC '02)*, pages 380–388.

Chatterjee, N. and Mohan, S. (2007). Extraction-based single-document summarization using random indexing. In *Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on*, volume 2, pages 448–455. IEEE.

Chew, P., Bader, B., Kolda, T., and Abdelali, A. (2007). Cross-language information retrieval using parafac2. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD07)*, pages 143–152.

Chiarello, C., Burgess, C., Richards, L., and Pollock, A. (1990). Semantic and associative priming in the cerebral hemispheres : Some words do, some words don't . . . sometimes, some places. *Brain and Language*, 38 :75–104.

Chomsky, N. (1957). *Syntactic Structures*. The Hague/Paris :Mouton.

Church, K. (1995). One term or two ? In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 310–318.

Church, K. and Hanks, P. (1989). Word association norms, mutual information and lexicography. In *Proceedings of the 27th Annual Conference of the Association of Computational Linguistics*, pages 76–83.

Coecke, B., Sadrzadeh, M., and Clark, S. (2010). Mathematical foundations for a compositional distributional model of meaning. CoRR. http ://dblp.uni-trier.de/rec/bib/journals/corr/abs-1003-4394.

Collins, A. and Quillian, R. (1969). Retrieval time from semantic memory. In *Journal of Verbal Learning and Verbal Behavior*, volume 8, pages 240–247.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing : Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning. ACM*, pages 160–167.

Dagan, I., Lee, L., and Pereira, F. C. N. (1999). Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34 :1–3,4–69.

Dang, H. (2006). Overview of duc 2006. In *HLT-NAACL*. Document Understanding Workshop.

de Loupy, C., Guégan, M., Ayache, C., Seng, S., and Moreno, J.-M. T. (2010). A french human reference corpus for multi-document summarization and sentence compression. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Deerwester, S., Dumais, S., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6) :391–407.

Erkan, G. and Radev, D. R. (2004). Lexrank : Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.(JAIR)*, 22(1) :457–479.

Fellbaum, C. (1998). *WordNet : An Electronic Lexical Database*. Bradford Books.

Firth, J. (1957). A synopsis of linguistic theory 1930–1955. *Studies in linguistic analysis*, Special volume of the Philological Society.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611.

Gentner, D. (1983). Structure-mapping : A theoretical framework for analogy. *Cognitive Science*, 7(2) :155–170.

Giles, J. (2005). Internet encyclopedias go head to head. *Nature*, 438 :900–901.

Goldstein, J. and Carbonell, J. (1998). Summarization : (1) using mmr for diversity - based reranking and (2) evaluating summaries. In *Proceedings of a Workshop on Held at Baltimore, Maryland : October 13-15, 1998*, TIPSTER '98, pages 181–195, Stroudsburg, PA, USA. Association for Computational Linguistics.

Golub, G. H. and Van Loan, C. F. (1996). Matrix computations (third edition). Johns Hopkins University Press, Baltimore, MD.

Gorman, J. and Curran, J. R. (2006). Random indexing using statistical weight functions. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 457–464, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gottron, T., Anderka, M., and Stein, B. (2011). Insights into explicit semantic analysis. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1961–1964. ACM.

Grefenstette, E. and Sadrzadeh, M. (2011). Experimental support for a categorical compositional distributional model of meaning. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Hadj Taieb, M. A., Ben Aouicha, M., and Ben Hamadou, A. (2013). Computing semantic relatedness using wikipedia features. *Knowledge-Based Systems*, 50 :260–278.

Harris, Z. (1954). Distributional structure. *Word*, 10(3) :146–162.

Higgins, D. and Burstein, J. (2007). Sentence similarity measures for essay coherence. In *Proceedings of the 7th International Workshop on Computational Semantics*, pages 1–12.

Hirao, T., Okumura, M., and Isozaki, H. (2005). Kernel-based approach for automatic evaluation of natural language generation technologies : Application to automatic summarization. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 145–152. Association for Computational Linguistics.

Hirst, G. and St-Onge, D. (1998). *FFellbaum, C. (Ed.), WordNet : An Electronic Lexical Database*, chapter Lexical chains as representations of context for the detection and correction of malapropisms., pages 305–332.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval (SIGIR '99)*, pages 50–57.

Hovy, E., Lin, C.-Y., Zhou, L., and Fukumoto, J. (2006). Automated summarization evaluation with basic elements. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.

Inderjeet, M. (2001). *Automatic Summarization*. John Benjamins Publishing.

Jaap, K. and Marijn, K. (2009). Is wikipedia link structure different ? In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 232–241.

Jarmasz, M. and Szpakowicz, S. (2003). Roget's thesaurus and semantic similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03)*, pages 212–219.

Ji, H., Ploux, S., and Wehrli, E. (2003). Lexical knowledge representation with contexonyms. In *Proceedings of the 9th Machine Translation Summit*, pages 194–201.

Jones, W. P. and Furnas, G. W. (1987). Pictures of relevance : A geometric analysis of similarity measures. *Journal of the American Society for Information Science*, 38 :420–442.

Kanerva, P. (1988). *Sparse distributed memory*. MIT Press.

Kanerva, P. (1993). *Sparse distributed memory and related models*. Oxford University Press, New York, NY.

Kanerva, P., Kristofersson, J., and Holst, A. (2000). Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd annual conference of the cognitive science society*, volume 1036. Erlbaum.

Karlgren, J. and Sahlgren, M. (2001). From words to understanding. In Uesaka, Y. and Kanerva, P. and Asoh, H. (Eds.). Foundations of Real-World Intelligence.

Ko, Y., Park, J., and Seo, J. (2002). Automatic text categorization using the importance of sentences. In *COLING*.

Kolda, T. and Bader, B. (2009). Tensor decompositions and applications. *SIAM Review*, 51 :455–500.

Landauer, T. and Dumais, S. T. (1997). A solution to plato's problem : The latent semantic analysis theory of the acquisition, induction and representation of knowledge. In *Psychological Review*, volume 104, pages 211–240.

Landauer, T., Foltz, P., and Laham, D. (1998). Introduction to latent semantic analysis. In *Discourse processes*, volume 25, pages 259–284.

Leacock, C. and Chodrow, M. (1998). *Fellbaum, C. (Ed.), WordNet : An Electronic Lexical Database. MIT Press.*, chapter Combining local context and WordNet similarity for word sense identification.

Lebret, R. and Collobert, R. (2014). Word embeddings through hellinger pca. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–490.

Lee, D. D. and Seung, H. S. (1999a). Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401 :788–791.

Lee, D. D. and Seung, H. S. (1999b). Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401 :788–791.

Lemaire, B. and Denhière, G. (2006). Effects of high-order co-occurrences on word semantic similarities. *Current Psychology Letters : Behaviour, Brain and Cognition*, 18.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorizationn. In *Advances in neural information processing systems*.

Li, Y., McLean, D., Bandar, Z. A., O'shea, J. D., and Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *Knowledge and Data Engineering, IEEE Transactions on*, 18(8) :1138–1150.

Lin, C.-Y. (2004). Rouge : a package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, pages 25–26.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*, pages 768–774.

Lin, D. and Pantel, P. (2001). Dirt-discovery of inference rules from text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001*, pages 323–328.

Lowe, W. (2001). Towards a theory of semantic space. In *In Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society*, pages 576–581.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments and Computers*, 28 :203–208.

Magnus, P. (2009). On trusting wikipedia. *Episteme*, 6(1) :74–90.

Mei, Q., Guo, J., and Radev, D. R. (2010). Divrank : the interplay of prestige and diversity in information networks. In Rao, B., Krishnapuram, B., Tomkins, A., and Yang, Q., editors, *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pages 1009–1018. ACM.

Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F., and Lanamäki, A. (2015). "the sum of all human knowledge" : A systematic review of scholarly research on the content of wikipedia. In *Journal of the Association for Information Science and Technology*, volume 66, pages 219–245.

Mihalcea, R., Corley, C., and Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *IN AAAI'06*, pages 775–780.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013b). Efficient estimation of word representations in vector space.

Mnih, A. and Hinton, G. (2008). A scalable hierarchical distributed language model. In *Advances in neural information processing systems*, pages 1081–1088.

Mnih, A. and Kavukcuoglu, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in neural information processing systems*, volume 26.

Montague, R. (1974). *Formal Philosophy.* Yale University Press, New Haven, USA.

Morin, F. and Bengio, Y. (2005). Hierarchical probabilistic neural network language model. In *AISTATS'05*, pages 246–252.

Nakayama, K., Hara, T., and Nishio, S. (2008). Wikipedia link structure and text mining for semantic relation extraction towards a huge scale global web ontology. In *Proceedings of Semantic Search Workshop (SemSearch)*, pages 59–73.

Nakov, P. and Hearst, M. (2008). Solving relational similarity problems using theweb as a corpus. In *Proceedings of ACL-08 : HLT*, pages 452–460.

Nelson, D., McEvoy, C., and Schreiber, T. (2004). The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, and Computers*, 36(3) :402–407.

Nenkova, A. and Passonneau, R. (2004). Evaluating content selection in summarization : the pyramid method. In *NAACL-HLT*.

Neto, J. L., Freitas, A. A., and Kaestner, C. A. (2002). Automatic text summarization using a machine learning approach. In *Advances in Artificial Intelligence*, pages 205–215. Springer.

Neto, J. L., Santos, A. D., Kaestner, C. A., and Freitas, A. A. (2000). Generating text summaries through the relative importance of topics. In *Advances in Artificial Intelligence*, pages 300–309. Springer.

Niwa, Y. and Nitta, Y. (1994). Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of the 15th International Conference On Computational Linguistics*, pages 304–309.

Ogden, C. K. (1930). *Basic English : A General Introduction with Rules and Grammar*. Kegan Paul, Trench, Trubner and Co.

Padò, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33 :161–199.

Pantel, P. and Lin, D. (2002a). Discovering word senses from text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 613–619.

Pantel, P. and Lin, D. (2002b). Document clustering with committees. In *Proceedings of the 25th Annual International ACM SIGIR Conference*, pages 199–206.

Pennington, J., Socher, R., and Manning, C. (2014). Glove : Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (ENMLP2014)*, pages 1532–1543.

Pingali, P., K, R., and Varma, V. (2007). Iiit hyderabad at duc 2007. In *NAACL-HLT 2007*.

Ploux, S. (1997). Modélisation et traitement informatique de la synonymie. *Linguisticae Investigatione*, 21 :1–28.

Ploux, S. and Victorri, B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires informatisés des synonymes. *Traitement Automatique des Langues*, 39 :161–182.

Pustejovsky, J. (1998). The generative lexicon. MIT Press, Cambridge, Massachusetts.

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the 9th Machine Translation Summit*, pages 315–322.

Ravichandran, D., Pantel, P., and Hovy, E. (1979). Randomized algorithms and nlp : using locality sensitive hash function for high speed noun clustering. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL '05)*, pages 622–629.

Reed, J. W., Jiao, Y., Potok, T. E., Klump, B. A., Elmore, M. T., and Hurson, A. R. (2006). Tf-icf : A new term weighting scheme for clustering dynamic data streams. *Machine Learning and Applications, Fourth International Conference on*, 0 :258–263.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Rohde, D. L., Gonnerman, L. M., and Plaut, D. C. (2006). An improved model of semantic similarity based on lexical co-occurrence. *Communication of the ACM*, 8 :627–633.

Rosset, S., Galibert, O., Bernard, G., Bilinski, E., and G., A. (2008). The limsi participation to the qast track. In *Actes de Working Notes of CLEF 2008 Workshop*.

Rumelhart, D. E., Hinton, G. E., and Ronald, W. J. (1986). Learning representations by back-propagation errors. *Nature*, 323 :533–536.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., and Scheffczyk, J. (2006). *FrameNet II : Extended Theory and Practice*. International Computer Science Institute, Berkeley, California. Distributed with the FrameNet data.

Sahlgren, M. (2005a). An introduction to random indexing. In *Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering, TKE*, volume 5.

Sahlgren, M. (2005b). An introduction to random indexing. in proceedings of the methods and applications of semantic indexing.

Sahlgren, M. (2006). The word-space model : Using distributional analysis to represent syn- tagmatic and paradigmatic relations between words in high-dimensional vector spaces.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24 :513–523.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing.

Scholkopf, B., Smola, A. J., and Muller, K.-R. (1997). Kernel principal component analysis. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN-1997)*, pages 583–588.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27 :379–423,623–656.

Shirakawa, M., Nakayama, K., Hara, T., and Nishio, S. (2009). Concept vector extraction from wikipedia category network. In *Proceedings of the 3rd International Conference on Ubiquitous Information Management and Communication*, pages 71–79. ACM.

Singhal, A., Salton, G., Mitra, M., and Buckley, C. (1996). Document length normalization. *Information Processing and Management*, 32 :619–633.

Sjöbergh, J. (2007). Older versions of the rougeeval summarization evaluation system were easier to fool. *Inf. Process. Manage.*, 43(6) :1500–1505.

Socher, R., Huang, E. H., Pennin, J., Manning, C. D., and Ng, A. Y. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 24*, pages 801–809. Curran Associates, Inc.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28 :11–21.

Steyvers, Mark ad Tenenbaum, J. B. (2005). The large-scale structure of semantic networks : Statistical analyses and a model of semantic growth. *Cognitive Science*, 29 :41–78.

Strube, M. and Ponzetto, S. P. (2006). Wikirelate ! computing semantic relatedness using wikipedia. In *AAAI*, volume 6, pages 1419–1424.

Torres-Moreno, J.-M. (2011). *Résumé automatique de documents : une approche statistique.* Recherche d'information et Web. Hermès.

Toutanova, K., Brockett, C., Gamon, M., Jagarlamudi†, J., Suzuki, H., and Vanderwende, L. (2007). The pythy summarization system : Microsoft research at duc 2007. In *NAACL-HLT 2007*.

Turney, P. D. (2001). Mining the web for synonyms : Pmi-ir versus lsa on toefl. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-01)*, pages 491–502.

Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, 32 :379–416.

Turney, P. D. (2007). Empirical evaluation of four tensor decomposition algorithms. Technical Report ERB-1152.Tech.rep., Institute for Information Technology, National Research Council of Canada.

Turney, P. D. (2008). The latent relation mapping engine : Algorithm and experiment. *Journal of Artificial Intelligence Research*, 33 :615–655.

Turney, P. D., Littman, M. L., Bigham, J., and Shnayder, V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03)*, pages 482–489.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning : Vector space models of semantics. *J. Artif. Int. Res.*, 37(1) :141–188.

Van de Cruys, T. (2009). A non-negative tensor factorization model for selectional preference induction. In *Proceedings of the Workshop on Geometric Models for Natural Language Semantics (GEMS-09)*, pages 83–90.

Van Rijsbergen, C. J. (1979). Information retrieval. Butterworths.

Voss, J. (2005). Measuring wikipedia. In *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*.

Vozalis, E. and Margaritis, K. (2003). Manolis vozalis manolis vozalis [pdf] from googlecode.com analysis of recommender systems algorithms. In *Proceedings of the 6th Hellenic European Conference on Computer Mathematics and its Applications (HERCMA-2003)*.

Weaver, W. (1955). *Machine Translation of Languages : Fourteen Essays*, chapter Translation. Locke, W. and Booth, D. (Eds.), . MIT Press, Cambridge, MA.

Weeds, J., Weir, D., and McCarthy, D. (2004). Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

William, B. and Lindenstrauss, J. (1984). Extensions of lipschitz mappings into a hilbert space. In *Conference in Modern Analysis and Probability*.

Wittgenstein, L. (1953). Philosophical investigations. Blackwell. Translated by G.E.M. Anscombe.

Wong, S. K. M., Ziarko, W., and Wong, P. C. N. (1985). Generalized vector spaces model in information retrieval. In *SIGIR ACM*.